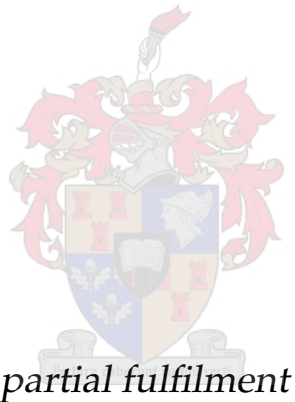
Defining and Assessing the Spectrum of Tuberculosis (TB) Disease: Application to Diagnosis and Prognosis

by

Tijani Ahmad Sulaimon



Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Science at Stellenbosch University

Supervisor: Prof. Alex Welte

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: November 12, 2020 .....

Abstract

Defining and Assessing the Spectrum of Tuberculosis (TB) Disease: Application to Diagnosis and Prognosis

Tijani Ahmad Sulaimon

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc

December 2020

Although Tuberculosis (TB) has been a topic of investigation for centuries, a satisfactory method of identifying patients with clinically active TB is yet to be discovered. This has two important negative consequences: Patients requiring treatment often remain untreated, and TB treatment is often administered 'presumptively' for patients, especially young children, who have symptoms suggestive of TB, even in the absence of laboratory test confirmation.

In this thesis, various aspects of the challenge of elucidating the spectrum of TB disease are explored: Key concepts in diagnosis, and defining performance metrics of diagnostics are explored; we consider the challenges of designing a study to evaluate a candidate diagnostic platform; simulated data, structurally aligned to a major ongoing study is analysed to demonstrate options for analysing and interpreting data in the early stages of TB marker evaluation; database schemas for the South African public sector TB treatment programme are explored, in particular relating the different handling of drug-sensitive TB (DS-TB) and drug-resistant TB (DR-TB); a preliminary analysis of DR-TB treatment outcomes is presented.

We demonstrate not only that sensitivity and specificity by themselves offer a limited view into the performance of a test, but also that these metrics are not intrinsic to a test, but vary with disease prevalence. More generally, the contextual 'spectrum of disease' – which, in practice, mainly means the distribution of the times since infection in the population being tested. In exploring study design options, we note the large number of assumptions that are required to fully specify details of conventional 'power calculations', suggesting that this is not a clear cut approach to choosing sample sizes. Given data generated by a well-understood biological process, we find that formal criteria driven by automated methods for optimising analysis, such as the least absolute shrinkage and selection operator (LASSO), provides little or no advantage over intuitively chosen diagnostic threshold criteria. A detailed mapping of data fields, linking the DS-TB and DR-TB treatment databases is produced, supporting consistent analysis across both types of TB, and facilitating analysis of some of the rich but complex structures in the DR-TB database. Although a significant number of drug-resistant TB patients do not have a recorded treatment outcome, unfavourable treatment outcomes, such as death, are found to be alarmingly common, and significantly associated with HIV status, history of previous TB treatment, age, and resistance patterns.

Better management of TB, a persistent and complex infectious disease, will require substantial additional research to be conducted in the coming years. It is hoped that this thesis will provide a meaningful resource to workers in this field, assisting them with numerous aspects of the search for better characterisation of the spectrum of TB; in particular, ways to access and analyse critical data and optimally design data gathering and analysis.

Uittreksel

Definisie en evaluasie van die spektrum van Tuberkulose (TB)-siekte: Toepassing tot diagnose en prognose

Tijani Ahmad Sulaimon

Departement Wiskuudige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MSc

Desember 2020

Alhoewel Tuberkulose (TB) al vir eeue die onderwerp van navorsing is, is daar steeds 'n gebrek aan 'n voldoende metode om pasiënte met kliniese aktiewe TB te identifiseer. Die laasgenoemde het twee belangrike negatiewe gevolge: Dikwels bly pasiënte wat dit benodig sonder behandeling, en dikwels word TB behandeling begin as 'n voorsorg maatreël, veral onder kinders, ten spyte van die gebrek van bewyse van aktiewe TB.

In hierdie tesis word verskillende aspekte van die uitdaging om die spektrum van TB-siekte toe te lig, ondersoek: Vernaamste konsepte in diagnose, en die definisie van prestasiemetings vir diagnostiese toetse is ondersoek; ons beskou die uitdagings om 'n studie te beplan om 'n kandidaat diagnostiese platform te evalueer; gesimuleerde data, met 'n gelyksoortige struktuur as 'n groot voortdurende studie, is ontleed om opsies vir analise en interpretasie van data voortgebring in die begin stadiums van TB-merker evaluasie, te demonstreer; databasis skemas vir die Suid-Afrikaanse openbare sektor se TB-behandelingsprogram word ondersoek, spesifiek om die verskillende bestuur van medikasie-sensitiewe TB (DS-TB) en medikasie-weerstandige TB (DR-TB) in verband te bring; 'n voorlopige ontleding van

DR-TB-behandelingsuitkomste word aangebied.

Ons demonstreer nie net dat sensitiwiteit en spesifisiteit alleenlik ‘n beperkte oorsig van die verrigting van ‘n toets voorstel nie, maar ook dat hierdie maatstawwe nie ’n intrinsieke deel van ’n toets is nie, maar dat dit varieer met die voorkoms van siektes. In die algemeen, die kontekstuele ‘spektrum van siekte’ – wat in praktyk hoofsaaklik die verspreiding van tye vanaf infeksie in die beproefde populasie beteken. By die ondersoek van opsies vir studieontwerp, neem ons kennis van die groot aantal aannames wat benodig word om die besonderhede ten volle te spesifiseer vir konvensionele onderskeidingvermoë analises, wat daarop dui dat dit nie ’n duidelike benadering tot die keuse van steekproefgroottes is nie.

Gegewe data wat gegenereer word deur ’n goed verstaanbare biologiese proses, vind ons dat formele kriteria gedryf deur outomatiese metodes vir die optimalisering van analise, soos die minste absolute krimp- en seleksieoperateur (LASSO), min of geen voordeel bied bo intuïtief gekose diagnostiese drempelkriteria. Die DS-TB en DR-TB behandelingdatabasis word gekoppel deur ‘n gedetailleerde belyning van die datavelde, wat die konsekwente analise van albei tipe TB ondersteun en die analise van sommige van die ryk maar ingewikkelde strukture in die DR-TB databasis fasiliteer. Alhoewel ‘n beduidende aantal pasiënte met medikasie-weerstandige TB geen aangetekende uitkoms van behandeling het nie, word daar gevind dat ongunstige behandelingsuitkomste, soos die dood, onrusbarend algemeen voorkom en dat dit sterk verband hou met MIV-status, geskiedenis van vorige TB-behandeling, ouderdom en patrone in weerstandigheid.

In komende jare sal aansienlike bykomende navorsing gedoen moet word om TB, ‘n volhardende en ingewikkelde oordraagbare siekte, beter te bestuur. Daar word gehoop dat hierdie tesis ’n sinvolle hulpbron aan werkers op hierdie gebied sal bied en hulle sal help met talle aspekte van die soeke na beter karakterisering van die spektrum van TB; veral maniere om kritiese data te bekom en te ontleed, en om data-insameling en -ontleding optimaal te ontwerp.

Acknowledgements

I thank SACEMA for funding this project, Professor Alex Welte for supervision, Professor Juliet Pulliam for support and advice, and Dr Muhammad Osman for helpful comments

I want to thank Dr Larisse Bolton, Cari van Schalkwyk, Jeremy Bingham, Dr Florian Marx, James Azam, Dr Elisha Are, and other SACEMA staff and students for useful feedbacks.

Dedications

To my parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aim and objectives

The thesis aims to better understand the spectrum of tuberculosis (TB) disease, analytical methods for creating useful categories out of complex multidimensional disease data, and ways of using these categories to understand and improve patient outcomes. Under this aim, we plan to: (1) Provide a coherent review of the challenges of TB and important concepts in diagnosis and treatment of TB; (2) Provide a theoretical review of the methods for evaluating the performance of a diagnostic marker; (3) Demonstrate how to analyse the output of a potential new marker, which closely reflects an actual TB marker project that is currently ongoing; (4) Provide a useful exploration of the TB treatment data in South Africa given that there are structural differences between how drug-sensitive and drug-resistant TB data are recorded; and (5) Analyse the drug-resistant TB treatment outcome.

1.2 Overview

Tuberculosis (TB) is a disease that has been known and continuously studied for hundreds of years, yet it still poses a significant threat to human lives. The disease remains a significant global health challenge, causing an estimated 1.5 million deaths annually [1]. It is the leading cause of death in South Africa, with an estimated 124 000 deaths in 2016 [2].

The main challenges facing TB treatment and eradication programs remain

diagnosis and treatment. TB diagnosis is particularly tricky, and despite advances in TB diagnostic research, a satisfactory test to rapidly and accurately diagnose TB remains a gap. The best available TB diagnostic test is the culture test, but it is expensive and requires an advanced laboratory. More importantly, it takes about 4 to 6 weeks to yield results. As a result, treatment decisions are made based on clinical symptoms, before culture test results are available. TB treatment has a long course (minimum of 6 months to 24 months) and often associated with unpleasant side effects and suboptimal outcomes. Reliable diagnostic tests that can rapidly diagnose TB, together with more efficacious antituberculosis drugs, are urgently needed to improve treatment outcome of TB patients [3].

This thesis investigates aspects of how to define the 'spectrum' of TB disease in ways that facilitate the investigation of diagnostic and prognostic indicators. In chapter 2, we review key concepts relevant to the investigation, including providing some more background on the burden of TB; exploration of mathematical ideas relevant to diagnostic performance; the particular challenges of characterising various levels of TB disease; and the current state of TB treatment options. Chapter 3 deals with important epidemiological and statistical aspects of designing a diagnostic accuracy study, including the strengths and limitations of various study designs; hypothesis formulation, particularly for hypothesis-driven studies; and various methods of calculating sample size. The challenges of evaluating a newly proposed TB marker as a diagnostic tool for childhood tuberculosis are demonstrated in Chapter 4. The challenges of managing drug-sensitive and drug-resistant TB treatment register data in South Africa are discussed in Chapter 5. For Chapter 6, a preliminary analysis of the drug-resistant TB database was conducted; providing specific descriptive analysis based on predictors such as age, sex, location, site of disease, previous treatment history, resistance pattern, and drugs used; and providing estimates of treatment outcomes (especially mortality) rates by various plausible predictors.

Chapter 2

Review

2.1 The Problem with TB

2.1.1 Burden of TB

TB in humans is caused by the bacterium known as Mycobacterium tuberculosis (Mtb). TB is transmitted from an infectious person to a susceptible person by breathing in air droplets released by the infectious person through coughing, sneezing, or talking [4]. TB is one of the leading causes of death globally [1, 5]. It is endemic in low and middle-income countries, contributing over 95% of World TB cases and deaths [1].

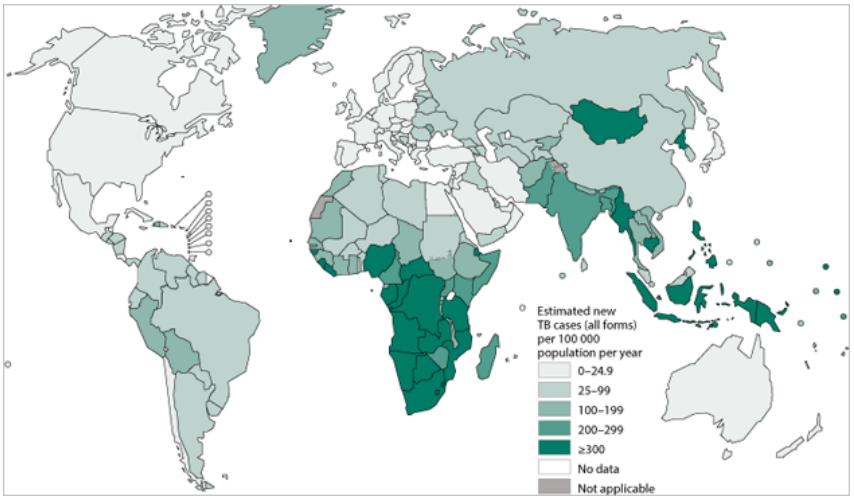


Figure 2.1: Estimated incidence of TB by countries. Reproduced from Glaziou et al. [6, p. 32].

In 2018, the incidence of TB was about 10 million, causing an estimated 1.5 million deaths [1]. Of those people who died of TB, 17% were co-infected with HIV [1]. South Africa, one of the TB endemic countries, reported about 124 000 TB deaths in 2016. Approximately 80% of those who died were co-infected with HIV [2].

2.1.2 Transmission, Stages, and Forms of TB

TB is not easily transmitted like other respiratory diseases [4]. The chance of an exposed person getting infected with TB depends on the:

- Number of TB bacteria released by an infectious person with active pulmonary TB through coughing, sneezing or talking;
- Volume of the air; and
- Duration of exposure to the bacteria [4, 7].

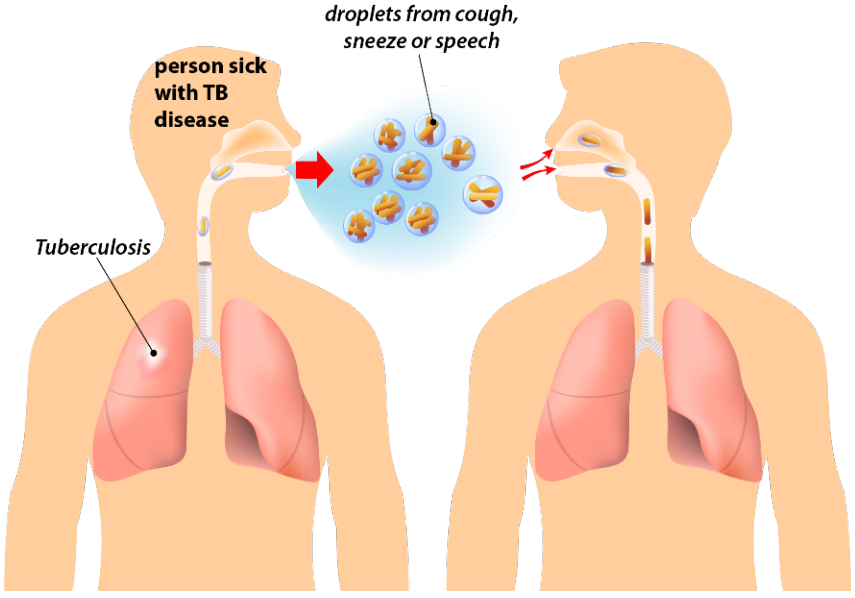


Figure 2.2: Transmission of TB. It is reproduced from Zhabska [8].

Droplets of Mtb are easily killed when exposed to direct sunlight. However, they survive longer in dark, indoor, and spaces with poor ventilation [7]. When droplets of Mtb are inhaled by a susceptible individual, they usually settle in the lungs where they slowly grow [3]. At this stage, the person is said to have latent TB or TB infection. People with latent TB are

asymptomatic; they do not become sick, and they are not capable of transmitting the infection. According to WHO [1], one-third of the population in the world has TB infection. TB disease, also known as active TB, occurs when a person with latent TB becomes sick. Only about $5-10\%$ of people with TB infection, who are not initiated on treatment, develop TB disease [5]. People with a weak or compromised immune system, such as children and people co-infected with HIV, are more likely to develop TB disease [3]. A person living with both HIV and TB infection has about $50-60\%$ chance of developing TB disease in a very short period [7].

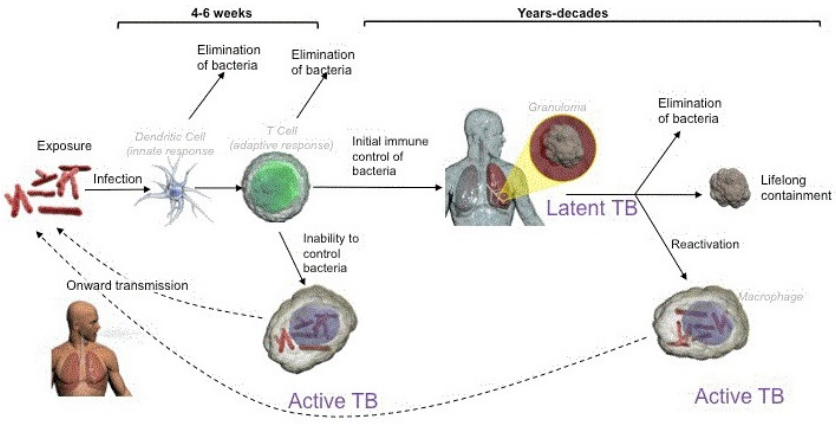


Figure 2.3: Natural history of TB infection. It is reproduced from Walsh [9].

Unlike HIV, which is distributed throughout the bloodstream, TB bacteria live in a specific site of the body. TB is classified as pulmonary TB (PTB) or extrapulmonary (EPTB) TB based on the site affected by the disease. TB that affects the lungs is called PTB, while TB that affects other parts of the body is called EPTB. PTB is the most common form of TB, occurring in about 80% of all TB cases worldwide [7]. In children, EPTB contributes to $30-40\%$ of TB cases [1].

2.1.3 Symptoms of TB

The general symptoms presented by people with active TB include fever, weakness, weight loss, lack of appetite, and night sweats. Additional symptoms of PTB include severe cough more than 20 days, coughing up phlegm and chest pain [7, 5]. Other symptoms of EPTB depend on the part of the

body affected by the bacteria, see Table 2.1. Whereas, people with TB infection do not show symptoms.

Table 2.1: Symptoms of Extra Pulmonary TB [7, 5].

Type of EPTB	Symptoms
Pleural TB	• unremitted cough • pleuritic chest pain • malaise
Lymph node TB	• swollen lymph node in the neck area or groin
TB meningitis	• vomiting • stiffness of the neck • seizures • nausea and drowsiness which may result in coma
Gastrointestinal TB	• diarrhoea • bleeding from the anus
Abdominal TB	• abdominal pain • abdominal swelling
TB of the kidney	• blood in the urine
Genitourinary TB	• irregular menstruation, pelvic pain, and infertility in women • slightly painful or painless lump inside scrotum in men

2.1.4 The Special Challenges of Young TB Patients in South Africa

According to WHO [1], childhood TB contributes about 10% of the total TB burden worldwide. In 2018, WHO [1] reported an estimated 1.1 million new cases of TB in children less than 15 years, with over 200 000 deaths globally. In South Africa, childhood TB contributes about 20% of all TB cases [2, 10]. About $30-40\%$ of TB in children is extrapulmonary; the rest are pulmonary [10].

Children have a higher risk of progressing from TB infection to TB disease. This progression happens in about one year, which may even be shorter in infants and very young children (< 4 years old) [10]. Severe and disseminated TB is more common in infants, associated with high morbidity and mortality [10, 11].

The diagnosis of TB in children is challenging due to acute presentation. Besides, diagnostic test samples, including sputum, are difficult to obtain from children as they require invasive procedures [12]. Even when the samples

are collected, TB bacteria present in them are often lower than the detection limit of the diagnostic test to be used [3].

2.2 Diagnosing TB

TB diagnosis refers to the identification of an active TB case that is, a patient with the disease caused by Mycobacterium tuberculosis (Mtb). Even though TB has been continuously studied for centuries and affects many people, yet we still don't have a satisfactory way of identifying patients with clinically active TB, and there are at least two consequences of this problem. First, clinicians fail to treat people who should be treated for TB due to false-negative test results. Also, precautionary TB treatment is administered, especially to young children, often in the absence of evidence of TB. We would prefer not to treat people unnecessarily because anti-tuberculosis treatment takes at least six months with numerous side effects [3]. It is also costly to the person and the health system.

In most TB endemic countries, TB diagnosis relies predominantly on microbiologic tests, including sputum smear microscopy [12]. These tests require invasive specimens such as sputum samples, which are difficult to obtain particularly in young children. In addition, diagnosis of TB with sputum samples has low value in patients with EPTB and patients co-infected with HIV [3].

2.2.1 TB Diagnostic Tests

Currently used TB diagnostic tests perform poorly in TB endemic countries where they are most needed. Sputum-smear microscopy has been the most used TB diagnostic test for more than a century. In countries such as South Africa, with high rates of TB infection, sputum-smear microscopy is often the first step in diagnostic procedures. It is a simple, inexpensive, and quick test that involves the identification of acid-fast bacilli (AFB) in sputum samples through a microscope. The infectivity of a patient is measured by the number of AFB seen [4]. It is an important test because it detects the most infectious active TB cases. However, the test is inaccurate; it misses above 40% of TB cases [7, 11]. The detection rate of active cases may even be lower

in countries with a high prevalence of HIV infection. This is because the detection of TB disease in a sample requires about 10 000 TB bacilli per ml of sputum [7], which children and many people with HIV and TB co-infection may not be able to produce [3]. It is not helpful in the diagnosis of EPTB.

TB culture test has also been used for TB diagnosis for over a century. It involves the identification of Mtb by growing it in a medium. The culture test is accurate; however, it is expensive and prone to contamination. It takes about $4-6$ weeks to yield results due to the slow growth of Mtb, leading to treatment decision delay. The important effect of treatment delay includes:

- increased severity especially in young children,
- continuous spreading of Mtb,
- death in patients with suppressed immune systems [3].

Polymerase chain reaction (PCR) based Xpert Mtb is a molecular test that can rapidly diagnose active TB in about two hours [12]. Even though it has enhanced TB diagnosis, it is still far from the ideal test for TB disease [3]. It detects more cases than Smear microscopy but less than the culture test. Xpert can differentiate Mtb from other bacteria. However, it cannot differentiate between dead and live Mtb [7]. The test is expensive and requires a constant supply of electricity. The cost of one Xpert test is about US $ 100 [3]. It is of limited value in children and patients with advance HIV.

Other tests commonly used include the Tuberculin Skin Test (TST) and chest radiograph. TST is a fast and straightforward TB diagnostic test that looks for Mtb [11]. Positive TST results indicate latent TB, which is essential in children as they develop TB disease faster [10]. A negative TST result does not necessarily mean the absence of TB bacteria. A patient infected with different bacteria other than Mtb may yield a false-positive TST result. Various studies have shown that vaccination with Bacilli Calmette-Guerin (BCG) may influence TST results [4].

Chest X-ray is a quick and convenient test that relies on the effect of TB bacteria on the person suspected of having TB [3, 11]. It can be used together

with TST to confirm or rule out active pulmonary TB in people with positive TST results [7]. The changes seen on a chest X-ray might suggest active TB; however, these changes are not specific to TB. This may lead to incorrect diagnosis as many diseases mimic TB disease on chest X-rays [3, 11]. A chest X-ray is essential in a patient who is unable to produce sputum for a bacteriological test or a patient who is HIV positive but has a negative Xpert result.

2.2.2 Gold Standard for Diagnosing TB

An ideal gold standard test is fast, easy, and cheap, with no spurious results. For a disease such as TB, there is no gold standard since there is no single test that can satisfactorily diagnose TB. Culture test and GeneXpert are currently the two best performing tests for diagnosing TB. The culture test is the closest we have to a gold standard in terms of accuracy, but it is prone to contamination and takes a long time to yield results. On the other hand, GeneXpert is fast but has low accuracy compared to culture. So, there is an urgent need for new point of care diagnostic assays that rely on non-invasive specimens [12]. For a new test to be evaluated, it needs to be validated against a gold standard or a reference standard. In this case, results from several available tests, in addition to clinical evaluation may be combined to form a reference standard.

2.3 Treating TB

TB is a preventable and curable disease [22]. The course of TB treatment has a duration of at least six months [3]. The treatment regimen contains multiple drugs to which the bacteria in a particular patient are susceptible. The success of treatment depends heavily on whether the drugs are taken in the right dosage for the prescribed interval. However, the unpleasant side effects of the drugs made it difficult for patients (especially very young children) to use them consistently [3]. Even though most of the bacteria are killed during the first eight weeks of treatment, if the treatment is not continued for a long duration, the surviving bacteria may cause the patients to become sick and infectious again. This may potentially lead to drug-resistance TB. TB is classified as either drug-sensitive TB (DS-TB) or

drug-resistant TB (DR-TB), based on whether the TB bacteria are sensitive or resistant to some TB drugs [7, 1].

Globally, the treatment success rate for new TB cases was about 85%, and only about 56% of people with DR-TB was successfully treated in 2017 [1]. DR-TB can be caused by:

- Poor patient management,
- Poor patient adherence to TB therapy, and
- Direct transmission from a sick person with DR-TB to a susceptible person [7].

About $70 - 80\%$ of all DR-TB cases in the world are caused by direct transmission [22].

2.3.1 Evaluating Success of TB Treatment

Monitoring and evaluation of TB treatment response is a crucial aspect of TB control programmes [7, 23, 24]. In practice, clinicians use patient's response to TB treatment to investigate treatment efficacy (in particular to investigate which patients are doing well or doing badly) over the course of treatment and to identify adverse reactions caused by the drugs used. Successful treatment of TB patients is key to eradicating TB. However, poor treatment outcomes can hinder the National TB Control Programme from reaching the End TB Strategy goals.

A robust investigation of treatment outcomes can help determine the vital factors associated with successful and poor treatment outcomes. The rate of treatment success can be improved when risk factors associated with poor treatment outcomes are systematically targeted. According to the SA National TB Management Guidelines, 2014 [7], TB treatment outcomes were grouped into seven categories summarised in Table 2.2.

Table 2.2: Definition of TB treatment outcomes (adopted from National Tuberculosis Management Guidelines 2014, Department of Health, South Africa [7]).

Treatment Outcome	**Definition**
Cured	A patient who was smear or culture positive at baseline and who is smear or culture negative in the last month of treatment and one previous occasion at least 30 days prior.
Treatment completed	A patient who was smear or culture positive at baseline and who has completed treatment but does not have a negative smear or culture in the last month of treatment and on at least one previous occasion at least 30 days prior.
Treatment success†	A patient who is cured or has completed treatment.
Died	A patient who died during treatment.
Treatment failure	A patient who was smear or culture positive at baseline and remains positive or converted to positive at month 5 or later during the course of treatment.
Treatment default	A patient whose treatment was interrupted for at least two consecutive months during treatment.
Moved out	A patient who is referred from a facility where the diagnosis and registration were made to another facility within the same district to continue treatment.
Transfer out	A patient who is referred to a facility in another and for whom treatment outcome is unknown.

Treatment success† is not a formal treatment outcome.

Chapter 3

Designing an Evaluation of a Diagnostic Marker

3.1 Diagnostic Concepts

The performance of a test, whose results may be binary or continuous variables, can be measured using different metrics, including sensitivity, specificity, and predicted values. The results obtained from the comparison of a diagnostic test to reference standard are usually summarised in a two by two contingency table described in Table 3.1.

Table 3.1: Two by Two contingency table indicating primary performance metrics for a diagnostic test.

	Target condition	
Diagnostic test	Present	Absent
Positive	TP	FP
Negative	FN	TN

TP is the number of the 'True Positives', those with the target condition that tested positive; FP is the number of the 'False Positives', those without condition that tested positive; FN is the number of the 'False Negative', those with the condition that tested negative; TN is the number of the 'True Negative', those without the condition that tested negative.

3.1.1 Sensitivity and Specificity

The two critical features of a good test are reproducibility and validity [13]. If a diagnostic test is reproducible, then its results should not change when used on the same patient or specimen repeatedly [14]. Validity is a measure of how well a diagnostic test can differentiate individuals with a target condition and those without the condition [13].

People often use the sensitivity and specificity of diagnostic tests to measure their performance. Sensitivity is the probability of a positive test result when the diagnostic test is used on an average person with the target condition. Specificity is the probability that the test result will be negative when used on an average person without the target condition. Sensitivity and specificity can only be utterly determined if the actual disease status of individuals is known. A perfect test is 100% sensitive and 100% specific with respect to the presence and absence of the target condition, respectively. That is, every patient with the target condition will test positive on the test. Alternatively, every patient without the target condition will test negative. In reality, there is no perfect test.

While some tests yield binary results (yes or no, positive or negative), most tests yield quantitative results on a continuous scale. The results of a test with quantitative results can be converted to binary results, using a threshold value, as shown in Figure 3.1.

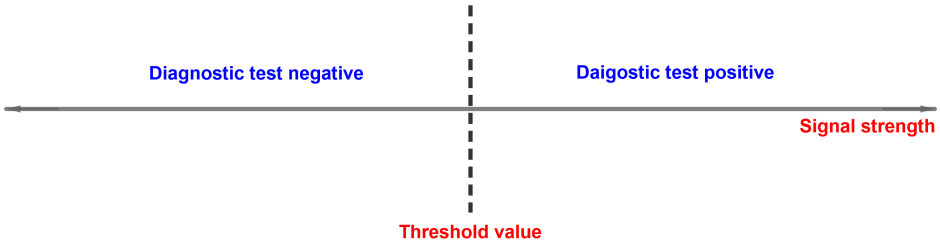


Figure 3.1: Converting numerical values on a continuous scale to binary, yes/no results.

As the threshold value is moved up the scale, patients will move from the diagnostic test positive to diagnostic test negative. In particular, there is a trade-off between sensitivity and specificity caused by varying threshold values. An increase in sensitivity leads to a decrease in specificity. Similarly,

a decrease in sensitivity leads to an increase in specificity[15].

The choice of sensitivity and specificity values of a diagnostic test depends on the clinical situation and the type of disease. A highly sensitive test is useful when false-negative results are to be avoided but at the expense of some false-positive results (classifying those without the condition as diseased). On the other hand, a highly specific test is useful when false-positive results are to be avoided but at the expense of classifying those with the target condition as non-diseased.

3.1.2 Sensitivity as a Function of Time Since Infection

The notion that sensitivity and specificity can each be summarised into a single number, robust across contexts, is limiting. Even though they do not directly depend on the prevalence of the disease, they do depend on other factors associated with prevalence [16]. More typically, the probability of successfully detecting an infection with a particular test is a function of time since infection (whether the specimen was collected two hours or one month after infection), and possibly numerous other additional factors, including age and co-infection with other pathogens.

At a group level, the sensitivity of a test will depend on the distribution of cases in the population on which the test is used, which in turn is dependent on the contextual distribution of times since infection. An infected individual has a curve like Figure 3.2, which depicts the probability of detecting disease at different times since infection.
The curve shows that test results immediately after an infection is negative but after some time, the test will almost certainly detect the infection. In a given population, different people take different time since infection to attain a particular value of probability, some get detected early, some get detected late, some never get detected as shown in Figure 3.3. Averaging over all the red curves gives the blue curve known as the sensitivity curve. It represents the probability, as a function of time after infection, that a randomly picked person would have a positive test result.

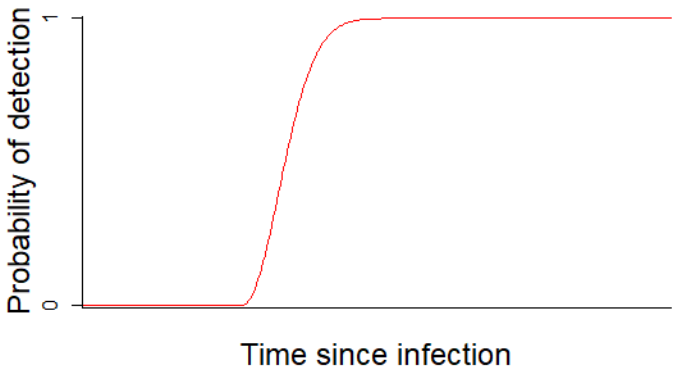


Figure 3.2: Probability that a specimen obtained at a particular time since onset of clinical TB (x-axis) and tested by a particular diagnostic procedure (implied by context) would result in a positive diagnosis. Note that this curve is intended to apply to a particular individual, who is experiencing a particular course of a disease.

Figure 3.3: Emergent test sensitivity as a function of time: Thin lines indicate detection probabilities of an ensemble of individuals. The bold line indicates the emergent net sensitivity - i.e. the probability, as a function of clinical time, that a randomly chosen subject will yield a positive test.

3.1.3 Context

Metrics for measuring diagnostic test performance are not robust across contexts, they vary in different populations. Context refers to the distribution of times since infection in a population. A context may be clinical (e.g. point of care, primary or secondary care) or population (e.g. survey or screening). A clinical context contains patients who are routinely presented

with similar characteristics (symptoms). So, it is essential to ask questions like what triggered the test? Why is the patient in the clinic? A population context contains a broader distribution of people. In particular, different contexts have a different spectrum of disease, which heavily depends on the prevalence and severity of disease (time since infection) [15].

For example, the performance of a particular test would be different when used in primary care and secondary care; this is because secondary care facilities deal with more advanced cases. Similarly, one cannot directly extrapolate the performance of a test in a hospital or clinical setting to a community because the number of cases and severity of disease would be higher in the hospital than in a general population. We would expect a test to perform terribly in the context of people who very recently got infected, i.e. times since infection is less than the diagnostic delay of the test.
If in a context, we know the density of times since infection in diseased people, then the sensitivity value in the context is given by:

$$\text{In-context sensitivity} = \int_0^\infty P(\tau)P_+(\tau)d\tau, \tag{3.1.1}$$

where $P_+(\tau)$ is the probability of getting a positive test result (blue curve in Figure 3.3), $P(\tau)$ is the density of time since infection, and τ is the time since infection.

3.1.4 Predictive Values

The problem with the sensitivity and specificity of a diagnostic test is that they tell us the probability of positive and negative test results respectively given true disease status, which we do not know [14, 16, 17]. In clinical practice, predicted values are more informative than sensitivity and specificity [15]. They tell us the probability of true disease status given test result, which we know. Positive predicted value (PPV) is the probability that the condition of interest is present given a positive test result. Negative predicted value (NPV) is the probability that the condition of interest is absent when the test result is negative.

A critical pitfall of predicted values is that they directly vary with the disease prevalence of the population where the test is used. Even if we assume

that sensitivity and specificity are robust properties of a test, across various contexts, the performance/interpretation of test results will vary with disease prevalence. For example, consider a diagnostic test which is 90% sensitive and 90% specific, used in a context with 10% prevalence. If a person chosen entirely at random from this population is tested and obtains a positive result, the probability of that person having the disease is 50%.

Table 3.2: Illustration of how to calculate predicted values and likelihood ratios of a hypothetical test used in a population of known prevalence.

	True		
Test	Positive	Negative	Total
Positive	90	90	180
Negative	10	810	820
Total	100	900	1 000

$$\text{Positive predicted value} = \frac{\text{True positive}}{\text{Test positive}} = \frac{90}{180} = 0.500,$$
$$\text{Negative predicted value} = \frac{\text{True negative}}{\text{Test negative}} = \frac{810}{820} = 0.988,$$
$$\text{Positive likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{specificity}} = \frac{0.9}{1 - 0.9} = 9.00,$$
$$\text{Negative likelihood ratio} = \frac{1 - \text{Sensitivity}}{\text{specificity}} = \frac{1 - 0.9}{0.9} = 0.111.$$

In Table 3.2, we considered the expected outcomes if we apply the test to 1 000 individuals, of whom 100 have the target condition. All we know about a person who just tested positive is that they are in the test positive row. This row contains 180 people, of whom 90 (half) are in the true positive column. Hence, PPV is $\frac{90}{180} = 50\%$. Similarly, if the test is used in population with 20%, the PPV is 69%. NPV is 99% and 97% respectively. This example illustrates that a low prevalence gives low PPV. The converse is also true. PPV and NPV can be written in terms of sensitivity, specificity, and prevalence as:

$$\text{PPV} = \frac{P \times \sigma}{P \times \sigma + (1 - P) \times (1 - \rho)} \tag{3.1.2}$$

$$\text{NPV} = \frac{(1-P)\times\rho}{(1-P)\times\rho + P\times(1-\sigma)} \tag{3.1.3}$$

where P is the prevalence, σ is sensitivity, and ρ is specificity.

3.1.5 Bayes Theorem and Likelihood Ratios

In clinical practice, diagnostic test results are not used in isolation; they are combined with clinical evaluation of patients. Before a test is performed, clinicians always have their suspicion or opinion about the presence or absence of the target condition in a patient, based on symptoms. This suspicion can be expressed as odds or probability (probability $= \frac{\text{odds}}{\text{odds}+1}$) of target condition being present before performing a test. This odds is known as pretest odds. When the test result is known, Bayes theorem can be used to combine test characteristics with patient characteristics (pretest odds) [18]. Bayes theorem for diagnostic test evaluation is given by:

$$\text{Post-test odds} = \text{Pre-test odds} \times \text{Likelihood ratios}, \tag{3.1.4}$$

where post-test odds are the odds of the target condition after knowing the test result. Likelihood ratios (LR) represents how likely is it that the target condition is present or absent given test result [18].

$$\begin{aligned}\text{Positive LR} = LR+ &= \frac{\begin{array}{c}\text{probability of test being positive}\\ \text{in a patient with the target condition}\end{array}}{\begin{array}{c}\text{probability of test being positive}\\ \text{in a patient without the condition}\end{array}}\\ &= \frac{\text{TP}}{\text{TP}+\text{FN}} \div \frac{\text{FP}}{\text{FP}+\text{TN}} = \frac{\text{sensitivity}}{\text{1 - specificity}}.\end{aligned}$$

The values of $LR+$ ranges from 0 to ∞. When $LR+$ is high, then in a given population, a positive test result is more likely in patients with the condition of interest. A perfect diagnostic test (sensitivity = specificity = 100%) has $LR+$ of $\frac{100\%}{1-100\%} = \infty$, while a useless diagnostic test has $LR+$ of 1, which means a positive test is as likely in patients with target condition as patients without the condition [16]. The negative likelihood ratio is defined as:

$$LR- = \frac{\begin{array}{c}\text{probability of test being negative}\\ \text{in a patient with the target condition}\end{array}}{\begin{array}{c}\text{probability of test being negative}\\ \text{in a patient without the condition}\end{array}}$$
$$= \frac{\text{FN}}{\text{TP}+\text{FN}} \div \frac{\text{TN}}{\text{TN}+\text{FN}} = \frac{\text{1 - sensitivity}}{\text{specificity}}.$$

The values of $LR-$ ranges from 0 to 1. It expresses how less likely a negative test is to be expected in patients with target condition as opposed to patients without the condition.

Consider some practical examples of Bayesian framework in equation 3.1.4. A test which is 95% sensitive and 95% specific has:

$$LR+ = \frac{0.95}{1-0.95} = 19 \quad \text{and} \quad LR- = \frac{1-0.95}{0.95} = 0.05.$$

Suppose Dr Gray examined a patient X based on symptoms, and estimated a pretest probability of 0.5, resulting in pretest odds of $\frac{0.5}{1-0.5} = 1$. Then the post-test odds for a positive test result is $1 \times 19 = 19$, giving a post-test probability of $\frac{19}{1+19} = 0.95$. This result will give Dr Gray greater confidence that patient X has the condition. Similarly, if the test result is negative, then the post-test odds is 0.05, giving a post-test probability of 0.05. This result will tell Dr Gray how likely is it to have a negative test if patient X has the disease, thus increasing his confidence that the condition is absent in patient X.

Now suppose Dr Gray has a very sick patient Y, with well-defined symptoms associated with the target condition, and he estimated pretest probability of 0.9. Then a positive test result will yield a post-test odds of 171, giving a post-test probability of 0.99. Hence, Dr Gray would even be more confident that patient Y has the condition. However, if the test result is negative, then we have a post-test odds of 0.45, giving a post-test probability of 0.31, which does not help to prove the absence of disease. These two scenarios again illustrate the importance of interpreting test results in the context of use.

3.1.6 The Idea of a Gold Standard

A gold standard diagnostic test is the best available test for determining the presence or absence of a target condition [19]. A perfect diagnostic test is a test with no wrong answers, i.e. no false-positive and false-negative results [15]. A gold standard test is imperfect, and nothing more than the best available test for diagnosing a condition at a particular time [20, 21].

In diagnostic research, gold standards are challenged continuously and replaced when appropriate because they are subject to continuous advancement [20]. Nevertheless, it is crucial to evaluate the performance of a novel test against a test that serves as the gold standard at a particular time, in a particular context [19, 15]. The performance metrics estimated for the new test are subject to reference test bias due to errors associated with the gold standard [19]. This problem has led to the development of various statistical methods for evaluating a novel test [15], with many unresolved issues and no universal solution [19].

3.2 Developing a New Diagnostic

TB and coronavirus infections are respiratory infections that affect mainly the lungs, common symptoms include cough and breathing difficulties [22, 25]. They are caused by microbes such as viruses and bacteria [1, 26]. In this section, we shall use the recent coronavirus to discuss some of the concepts in developing new diagnostics.

In December 2019, the WHO China received notification of multiple cases of severe pneumonia in Wuhan, China [25, 26]. Samples taken were sent to the Chinese Center for Disease Control and Prevention for investigation. The presence of a virus belonging to the coronavirus family was found in patient samples, through genome sequencing [26]. The results of known coronavirus tests that were performed suggest the virus is novel and was termed SARS-CoV-2, the causes of the disease COVID-19 [26].

Molecular tests, also known as polymerase chain reaction (PCR) tests, were the first set of tests designed to diagnose infection with COVID-19, and they

are the current gold standard. PCR tests rely on detecting RNA sequence consistent with the RNA sequence in the genome of the COVID-19 virus [27]. Even though it is very fast to design a molecular test, it can not be used as a rapid on-sight test because it requires samples to be shipped to a centralised laboratory [28].

Due to the high infectivity of COVID-19 rapid, accurate, and easy to use tests are urgently needed to detect and treat infected patients. Unlike molecular tests that rely on detecting RNA sequence, antigen tests rely on detecting the presence of viral proteins that are part of the COVID-19 virus [27]. They are often fast and easy to use, but they are not as accurate as compared to molecular tests.

Antibody tests rely on finding antibodies against the COVID-19 virus [27]. Antibodies are specific proteins induced by the immune system in response to a particular infection. The result of an antibody test may be used as a diagnostic test for COVID-19. Unfortunately, a positive test does not only mean the patient has the infection now, but it may also mean the patient might have had the infection in the past [27].

Now, suppose an antigen or antibody test is proposed, and we want to know the performance of the test in comparison with the reference standard, the PCR test, then a study has to be conducted. Such a study is known as a 'diagnostic performance study'.

Appropriate study design must be used to appraise the performance of a diagnostic test [17, 29, 30]. This design should be done in a population that is clinically relevant [30]. There are two basic designs of diagnostic test performance studies, namely the single-gate and the two-gate designs [29, 30].

The single-gate design, shown in panel A and B of figure 3.4, is also regarded as a diagnostic cohort study, and sometimes matched or paired design. In this type of design, the new test and the reference standard are used on the same set of study subjects with unknown disease status [30]. For example, two blood-based assays can be used on blood samples of each subject in the study. The single-gate design has only one inclusion criteria,

which is determined by the clinical presentation or symptoms of subjects [15]. It represents a situation where the test under investigation would be used in clinical practice [15].

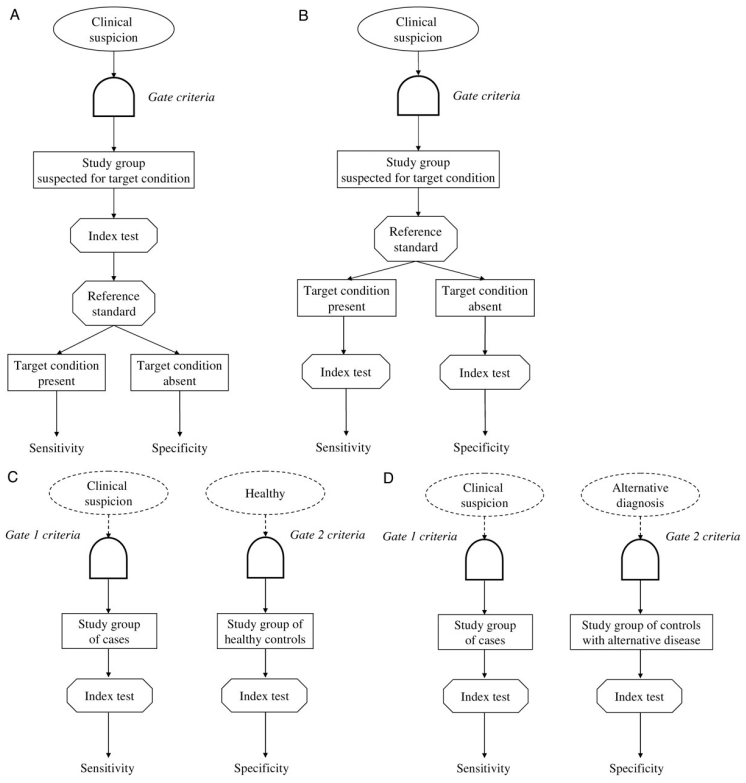


Figure 3.4: Study design for diagnostic accuracy study. Reproduced from Rutjes et al. [30].

The two-gate design, shown in panel C and D of figure 3.4 is similar to case-control study design in epidemiology. The former include subjects based on true disease status, giving two arms; one group with the condition of interest (case), and the other group without the condition (control). In contrast, the latter include subjects based on exposure [17, 30]. In two-gate design, sometimes regarded as unmatched, unpaired or parallel design, the new test is applied to the two arms of subjects after the reference standard has been used to group subjects into two arms.

The two-gate design is subject to spectrum bias because it is difficult to capture the full spectrum of health and disease status of the population in which the test would be used [15]. If the case arm contains subjects with

more advanced disease of interest, then we will overestimate how well the new test can identify subjects with the disease (sensitivity of the new test). Similarly, if control subjects (in the test performance study) have a low level of 'cross-reacting' antibodies, compared to the populations in which the test is going to be used, then the specificity (how well the new test can identify subjects without the disease) of the new diagnostic tends to be overestimated. Cross-reactivity is the tendency for a test to interact with something other than its intended target. Although in the early stage of developing a new diagnostic, the two-gate design may be useful to decide whether the test is of potential use or justifies serious investment [30].

3.3 Hypothesis Formulation

Diagnostic test performance studies may be hypothesis-driven or non-hypothesis-driven. Studies that are hypothesis-driven aims to answer a specific well-defined research question [31]. The hypothesis for testing the sensitivity/specificity (S_N) of a new diagnostic test against that of a reference standard (S_R) can be constructed in the following way:

1. Test for equality – to show that the difference in performance between a new test and a reference standard is statistically significant, we may consider the hypotheses in equation 3.3.1.

$$H_0 : S_N - S_R = 0 \quad \text{against} \quad H_1 : S_N - S_R \neq 0. \tag{3.3.1}$$

   We discard the null hypothesis if the data is unlikely to be produced by tests of equal sensitivity/specificity – the p-values measure this probability [32, 33]. The possible outcomes of the experiments are:

   a) The null hypothesis NOT rejected – This does not imply that the tests have equal performance; it means we lack statistical evidence of the difference.

   b) We reject the null hypothesis in favour of the alternative hypothesis that the new test performs better than the reference standard (this implies the test has satisfied the superiority test with $H_0 : S_N - S_R \geq 0$ and $H_1 : S_N - S_R > 0$). It means that if the new test is no better than the reference standard, then there is a

low probability (p-value is small) of the new test outperforming the reference standard by some margin greater than zero.

c) We reject the null hypothesis in favour of the alternative hypothesis that the reference standard performs better than the new test. It means that if the reference standard is no better than the new test, then there is a low probability (p-value is small) of the reference standard outperforming the new test by some margin greater than zero.

2. Test for strong superiority – to establish that the sensitivity/specificity of a new test is superior over that of the reference standard, we may use the hypotheses in equation 3.3.2.

$$H_0 : S_N - S_R \leq c_s \quad \text{against} \quad H_1 : S_N - S_R > c_s. \tag{3.3.2}$$

The rejection of the null hypothesis indicates the difference in sensitivity/specificity between the new test and the reference standard is greater than a clinically meaningful difference of c_s. Hence we say that the new test is more sensitive/specific than the reference standard [33].

3. Test for non-inferiority – to establish that the performance of the new test is NOT worse than that of the reference standard by more than a clinically significant margin (non-inferiority margin), we consider the hypotheses in equation 3.3.3.

$$H_0 : S_R - S_N \geq c_s \quad \text{against} \quad H_1 : S_R - S_N < c_s. \tag{3.3.3}$$

The idea is to discard the null hypothesis and conclude that the difference in performance between the new test and the reference standard is less than a clinically significant margin, c_s. Hence we say the new test is as sensitive/specific as the reference standard [33].

4. Test for equivalence – The hypotheses in equations 3.3.2 and 3.3.3 require prior information regarding the comparison of the new test to the reference standard, which in practice is mostly unavailable [33]. An alternative way to avoid such ambiguity is to formulate the hypotheses as described in equation 3.3.4.

$$H_0 : |S_N - S_R| \geq c_s \quad \text{against} \quad H_1 : |S_N - S_R| < c_s. \tag{3.3.4}$$

When H_0 is rejected, we conclude that the difference between the new test and the reference standard is clinically meaningless.

Studies that are not hypothesis-driven are in the form of a random search. However, they should at least have a clear goal or direction [31].

3.4 Sample Size Calculation

One of the critical factors that determine a reliably precise measure of an outcome of interest in a diagnostic performance study, usually sensitivity and specificity, is the sample size (number of participants in a study) [31]. While sparsity of participants in a study yields an imprecise or unreliable estimate of an outcome, too many participants than required may have little or no effect on observed outcome [31]. Thus, it is imperative to have an idea of the optimal sample size required to yield a precise estimate of the outcome. Sample size can be calculated based on power and precision [33], and sometimes based on cost or availability [34].

3.4.1 Sample Size for Power

In a comparative study, power calculation is done to avoid a type II error, when you fail to observe a difference when there is a difference [28]. Given a significant level and sample size, the power of a study is the probability of observing a difference if it truly exists.

To calculate sample size for sensitivity or specificity, based on power, we need the following information: (1) the expected sensitivity (or specificity) value; (2) the desired level of confidence, denoted by $(1 - \alpha)$, where α is the probability of observing a difference where there is no meaningful difference (type I error); (3) the desired power, denoted by $(1 - \beta)$, where β is the probability of not observing a difference when in fact there is a significant difference (type II error); and (4) the precision of estimates one is willing to achieve, considered here as the maximum marginal error [32, 33, 35].

Suppose we want to investigate whether there is a difference between the sensitivity (or specificity) of a test and a reference value, we may consider the following hypothesis;

$$H_0 : s = s_0 \quad \text{versus} \quad H_1 : s \neq s_0,$$

where s is the true sensitivity (or specificity) of the test and s_0 is the sensitivity (or specificity) under the null hypothesis. The approximate sample size needed to estimate s with $(1-\alpha)\%$ confidence and $(1-\beta)\%$ power is given by;

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{s_0(1-s_0)} + Z_\beta\sqrt{s(1-s)}\right]^2}{(s-s_0)^2}, \tag{3.4.1}$$

where $Z_{\frac{\alpha}{2}}$ and Z_β are the upper quantile values of the standard normal distribution at $\frac{\alpha}{2}$ and β, respectively [32, 33]. For example, a 95% confidence level and 80% power will have α and β values of 0.05 and 0.2, resulting in Z values of 1.96 and 0.84, respectively.

Similarly, to test whether there is a difference between the sensitivity (or specificity) of two diagnostic tests, we use the following hypothesis;

$$H_0 : s_1 = s_2 \quad \text{versus} \quad H_1 : s_1 \neq s_2.$$

The approximate sample size to achieve a power of $(1-\beta)\%$ with $(1-\alpha)\%$ confidence is given by;

$$n = \frac{\left[Z_{\frac{\alpha}{2}}\sqrt{2s(1-s)} + Z_\beta\sqrt{s_1(1-s_1)+s_2(1-s_2)}\right]^2}{(s_1-s_2)^2}, \tag{3.4.2}$$

where s_1 and s_2 are the expected sensitivity (or specificity) values of the two diagnostic tests, s is the average of s_1 and s_2 [32, 33].

Suppose we have a sample size, and we ask what power do we have to observe a particular difference? To answer this, we simply solve for Z_β from equation 3.4.1 and 3.4.2 as required and then calculate power $= 1-\beta$. For example, given 100 samples with target condition and actual sensitivity of 90%. We want to test the null hypothesis that the sensitivity is 70%. We discard the null hypothesis if we are sure that the sensitivity is higher than

70%, at a p-value of 0.05 or lower. What is the probability that we will correctly conclude that sensitivity is higher than 70%?

$$H_0 : s = 0.7 \quad \text{versus} \quad H_1 : s > 0.7 \quad \text{(one-sided test)}.$$

From equation 3.4.1,

$$1 - \beta = \Phi\left(\frac{(s - s_0)\sqrt{n} - Z_\alpha\sqrt{s_0(1 - s_0)}}{\sqrt{s(1 - s)}}\right)$$

$$= \Phi\left(\frac{(0.9 - 0.7)\sqrt{100} - 1.64 \times \sqrt{0.7(1 - 0.7)}}{\sqrt{0.9(1 - 0.9)}}\right)$$

$$= 0.9999837 \approx 1.$$

Hence we have approximately 100% probability of correctly concluding that the sensitivity is higher 70%. Now, suppose we have two tests with sensitivities 0.7 and 0.9. We wish to test the null hypothesis that their sensitivity is the same and discard the null if it appears that one is better than the other, at a p-value of 0.05 or lower. What is the probability that we will reject the null hypothesis in favour of the correct ordering of sensitivity?

$$H_0 : s_1 = s_2 \quad \text{versus} \quad H_1 : s_1 > s_2 \quad (s_1 = 0.9 \text{ and } s_2 = 0.7).$$

From equation 3.4.2,

$$1 - \beta = \Phi\left(\frac{(s_1 - s_2)\sqrt{n} - Z_\alpha\sqrt{2s(1 - s)}}{\sqrt{s_1(1 - s_1) + s_2(1 - s_2)}}\right)$$

$$= \Phi\left(\frac{(0.9 - 0.7)\sqrt{100} - 1.64 \times \sqrt{2 \times 0.8(1 - 0.8)}}{\sqrt{0.9(1 - 0.9) + 0.7(1 - 0.7)}}\right)$$

$$= 0.97.$$

3.4.2 Sample Size for Precision

Calculating sample size based on power has several complicated problems. The method is convoluted and requires significant numbers of assumptions or prior information. They also suffer the problem of dichotomising studies into statistically significant or not significant, powered or underpowered,

and informative and non-informative, using p-value, which has many unresolved issues [35].

An alternative way is to calculate sample size based on desired precision, to get away with problems associated with statistical hypothesis testing [31]. This method provides an easy way to examine how the sample size varies with a varying relative or absolute precision and width of the confidence interval. This relationship can be represented graphically by precision against plotting sample size. The graph in Figure 3.5 shows absolute precision obtained over a range of sample size, $100 - 500$, in an attempt to estimate sensitivity (or specificity).

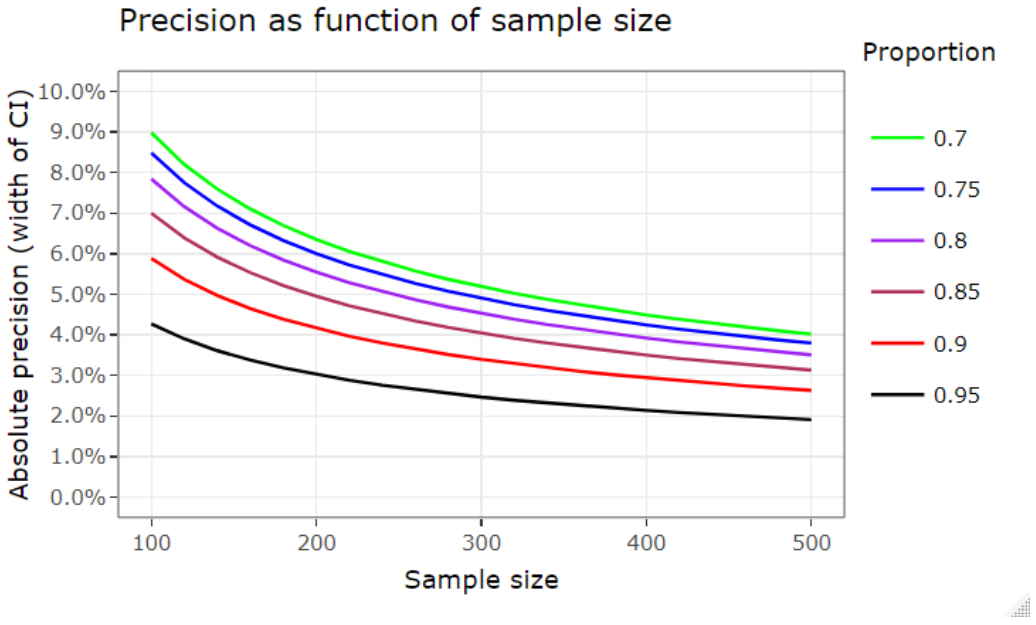


Figure 3.5: Calculating sample size for sensitivity/specificity based precision. CI, confidence interval.

3.4.3 Sample Size Based on Cost and Availability

Consider two proportions p_1 and p_2 with sample sizes n_1 and n_2, respectively. For an estimate of the difference in proportions, the precision, that is the width of the confidence interval, is $Z_{\frac{\alpha}{2}} \times \sigma_\Delta$. Where α is the confidence

level, and σ_Δ is the standard error. Since α is a constant, chosen arbitrarily, then the precision only depends on the standard error of the difference, given by

$$\sigma_\Delta = \sigma_{(\sigma_1 - \sigma_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

To maximise the precision at which the difference in proportions is measured, we need to minimise the standard error. Suppose the cost of each subject with and without a target condition is c_1 and c_2 respectively, then the total cost of obtaining the subjects in the two groups is $c = c_1 n_1 + c_2 n_2$. This leads to an optimisation problem stated as;

$$\begin{aligned} &\text{minimize } \sigma_\Delta(n_1, n_2) = \text{minimize } \left(\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right) \\ &\text{subject to} \quad c = c_1 n_1 + c_2 n_2 \end{aligned} \tag{3.4.3}$$

If we choose n_1 then $n_2 = \frac{c - c_1 n_1}{c_2}$. So equation 3.4.3 reduces to

$$\sigma_\Delta(n_1) = \left(\frac{p_1(1-p_1)}{n_1} + \frac{c_2 p_2(1-p_2)}{c - c_1 n_1} \right)^{\frac{1}{2}}$$

$$\begin{aligned} \frac{\partial \sigma_\Delta}{\partial n_1} &= -\frac{p_1(1-p_1)}{n_1^2} + \left[c_2 p_2(1-p_2) \times -\frac{1}{(c - c_1 n_1)^2} \times -c_1 \right] \\ &= -\frac{p_1(1-p_1)}{n_1^2} + \frac{c_1 c_2 p_2(1-p_2)}{(c - c_1 n_1)^2} = 0 \\ &\iff \frac{p_1(1-p_1)}{n_1^2} = \frac{c_1 c_2 p_2(1-p_2)}{(c - c_1 n_1)^2} \\ &\iff \frac{\sqrt{p_1(1-p_1)}}{n_1} = \frac{\sqrt{c_1 c_2} \times \sqrt{p_2(1-p_2)}}{c - c_1 n_1} \\ &\iff n_1 = \frac{c\sqrt{p_1(1-p_1)}}{c_1\sqrt{p_1(1-p_1)} + \sqrt{c_1 c_2}\sqrt{p_2(1-p_2)}}. \end{aligned}$$

Similarly, $n_2 = \dfrac{c\sqrt{p_2(1-p_2)}}{c_2\sqrt{p_2(1-p_2)} + \sqrt{c_1 c_2}\sqrt{p_1(1-p_1)}}$.

Hence, we choose n_1 and n_2 based on

$$\frac{n_2}{n_1} = \sqrt{\frac{p_2(1-p_2)c_1}{p_1(1-p_1)c_2}}. \tag{3.4.4}$$

Nam [36] and Cochran [37] derived sample size for two means with equal variance as a function of cost and availability, which was termed the 'square root rule' by Gail et al. [38]. In equation 3.4.4, we derived a formula for estimating the sample size of two proportions based on cost and availability.

3.4.4 The Importance of Blinding

Blinding is a vital aspect of diagnostic performance studies. The knowledge of the true disease status of subjects can impact the appraisal of the diagnostic test under study. So those performing the new test must be blinded from the true disease status and clinical information of subjects. Similarly, those performing the reference test should be unaware of the results from the new test [17]. To ensure measurements are reproducible, the examiners performing a test must be blinded from the previous results of the same subjects.

Chapter 4

The Proposed New TB Marker

4.1 Introduction

TB diagnosis is difficult, particularly among children. Previous studies at Desmond Tutu TB Centre (DTTC) have led to an availability of stored specimens taken from patients with particularly useful clinical data. A group currently based at Tulane University recently proposed an assaying platform for potentially detecting previously described TB-associated markers with new levels of sensitivity [39, 40]. In an ongoing collaboration, the new platform is being applied to some of these well-characterised specimens from DTTC, in particular, to measure the presence of 10-KDa culture filtrate protein (CFP-10). The marker, CFP-10, is a well-known protein secreted by active Mtb. Even though the protein is well known, no study has yet figured out a method to robustly measure it in a way that is adequate for clinical purposes [12, 40]. This chapter demonstrates some analyses that have been proposed to be applied to the data emerging from this project.

It was hoped, based on promising preliminary results, that the proposed assay for detecting some informative combination of these markers would not only serve as a useful clinical diagnostic tool but potentially also inform the assessment of 'treatment response'. Monitoring treatment response is essential because children are more likely to have paucibacillary TB disease (a low bacteria load of Mtb [5]), and it is increasingly evident that the bacteria are easily killed in the first few months of TB treatment course. We would like to know which patients do particularly well, or particularly severely, over

the course of treatment. This may potentially guide decisions to shorten the course of treatment in some patients, which is important because the standard treatment takes six months with unpleasant adverse reactions. The study aims to investigate various methods of detecting differences between the specimens from 'confirmed TB' cases and 'confirmed negatives'. Confirmed TB means a patient who is positive on culture or Xpert and confirmed negative is defined as a patient who has no well-defined TB symptoms and no bacteriological confirmation (culture and Xpert) of TB. In the main study, there are lots of specimens of probable TB, a patient with well-defined TB symptoms but no bacteriological confirmation of TB, but these will only be analysed in detail if at least there is a reasonable good naive sensitivity and specificity from the initial analysis.

4.2 The Platform

The proposed platform for monitoring and quantification of TB antigen from blood samples involves the use of 'multiple reaction monitoring' mass spectrometry. Mass spectrometry is an analytical technique used to measure the mass-to-charge ratio of ionised molecules [41]. Typically, analytical samples are vaporised and then ionised to give charged fragments, which can be accelerated to move on a straight line. An electromagnet is used to deflect accelerated ions off their path [41]. The quantity of force the charge feels from the electromagnet is proportional to its charge, and the path which the ion takes after deflection is as a result of its mass-to-charge ratio (m/z) [41]. The output of mass spectrometry is signal strength (relative abundance) as a function of fragment mass, produced by the reaction. A large target molecule is expected to break down reproducibly into a particular set of fragments which will contribute to several peaks in the spectrum.

Selective reaction monitoring (SRM) is a highly sensitive and specific analytical technique that can selectively identify and quantify a specific peptide present in complex mixtures of other peptides [42]. The SRM has two mass selection stages: the first mass selection stage selects the peptide ion of interest and then fragments it. The second mass selection stage selects a specific fragment ion for the detector, the instrument that records the mass-to-charge ratio of each fragment ion [42]. Multiple reaction monitoring (MRM)

is an application of SRM where more than one specific peptide is selected for the detector in the second mass selection stage. In Figure 4.1, serum samples are digested by trypsin, which breaks down proteins into small peptides. Target peptide is selected and then undergo fragmentation, followed by the selection of multiple specific peptide ions for the detector.

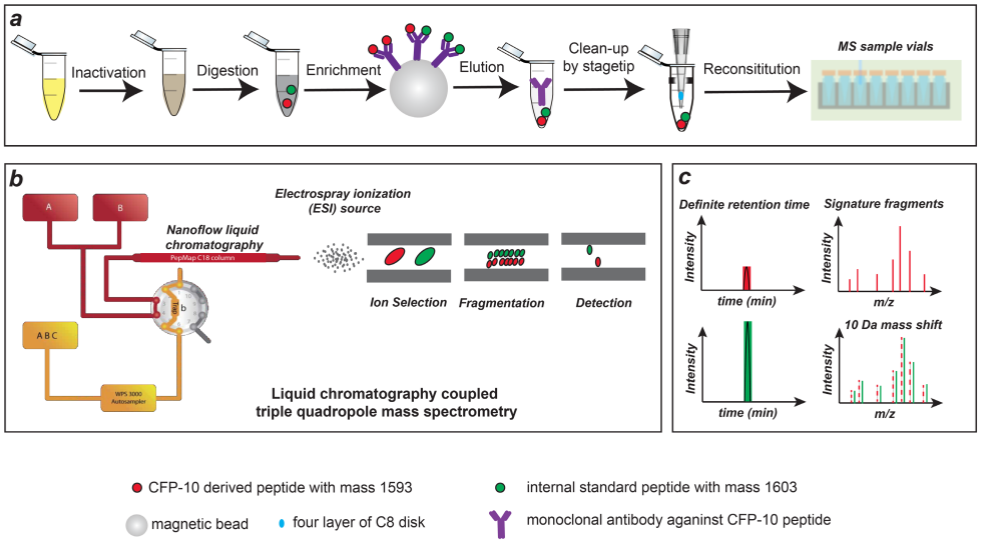


Figure 4.1: Workflow for sample detection using MRM: (a) The stages of sample preparation for mass spectrometry. Serum samples are trypsin digested and mixed with labelled internal standard to ensure accurate and reproducible quantification of target peptide [43]; (b) Recognition of peptide fragments of interest, the stage of the second mass selection; (c) Peptide quantification is derived by taking the ratio of mass spectra intensity of CFP-10 derived peptide and the labelled internal standard peptide.

4.3 The Data

The raw data from MS is a spectrum where the x-axis represents mass, and the y-axis represents signal strength, see part c of Figure 4.1. In a simple idealised situation where all the molecules subject to MS have precisely the same mass, there would be a single narrow peak in the spectrum. In practice, when MS is based on a clinically obtained specimen, there is a complex spectrum with multiple peaks of different sizes and potentially significant background noise.

By running standard specimens with explicitly controlled CFP-10 inputs, we can identify the peaks that are likely to arise from clinical specimens

with significant amounts of CFP10. The identification of peaks is based on a value between 0 and 1, regarded as 'rdotp'. The rdotp value is used to judge how similar are the observed peaks to the ones we expected to observe. When rdotp value is 1, the relative peak heights in the analytical specimen are precisely the same as the reference standard. The rdotp value is derived by taking the inner product of the peaks from the target analyte and the unknown specimen. Raw spectra are aggregated into peak areas and these can be reported by the system software directly. The final dataset is described in Table 4.1.

Table 4.1: Description of summary data from MS.

Column	Description
1–3	patient ID, sample ID, and analyst
4–5	date sample was run and sample volume
6	month sample was collected recorded as baseline, month 2 or month 6
7	TB status recorded as confirmed TB, unconfirmed TB, or ill control
8–12	individual peak intensity of internal standard peptide with mass 1603
13	sum of individual peak intensity of internal standard peptide with mass 1603
14	rdotp value of internal standard peptide with mass 1603
15–16	minimum start time and maximum end time
17–21	individual peak intensity of CFP-10 derived peptide with mass 1593
22	sum of individual peak intensity of CFP-10 derived peptide with mass 1593
23	rdotp value of CFP-10 derived peptide with mass 1593
24	ratio of CFP-10 derived peptide and internal standard peptide
25	CFP-10 status based on the proposed case definition recorded as either TB or Not TB

Confirmed TB, patients who are positive on culture or Xpert; unconfirmed TB, patients with well-defined TB symptoms but no bacteriological confirmation of TB; ill control, patients without well-defined TB symptoms and no bacteriological confirmation of TB; rdotp, a value between 0 and 1, used to judge how similar are the observed fragments to the ones we expected to observe; internal standard, a fixed amount of known substance added to every analytical sample for accurate and reproducible quantification.

Given the non-availability of a final dataset from the study, two hypothetical data was simulated. To date, specimens from 156 subjects at different time points have been sent to the laboratory; amongst these, 42 had confirmed TB disease and 114 controls. Once the final data has been produced from these specimens, several analyses will be done to decide whether this assaying platform justifies more serious investment. Using simulated data, we want to demonstrate how the data might look like, how to analyse it, and how it can be interpreted. The two simulated datasets have variables similar to the summary data described in Table 4.1. The first dataset was simulated such that the cases can be separated from the controls using two threshold values. The second dataset was simulated such that the cases and the controls are well mixed.

4.4 Analysis

All analyses were conducted using R statistical software. The first analysis relies on using the two-threshold criterion, the total peak area (sum of the individual peak intensity of derived peptide) and the corresponding rdotp value. The proposed case definition of positive CFP-10 status is defined as a total peak area greater than 200, and rdotp value greater than 0.5. The trade-off between sensitivity and specificity was investigated by varying the two threshold values.

For biological reasons, derived peptides are relatively abundant in different quantities of which, some are easy to measure, some are difficult to measure. So, instead of naively taking the sum of individual peak intensity, we may take a linear combination (weighted peak area). One way of doing this is to use the least absolute shrinkage and selection operator (LASSO) method. The LASSO method is a statistical model that aims to select one or a few predictors that strongly explains the response variable [44]. It provides an easy way to reduce the dimensionality of massive data.

Consider the logistic regression model,

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta X, \tag{4.4.1}$$

where $p(X)$ is the conditional probability $p(Y = \text{“1”}|X)$. The logistic regression aims to estimate the parameter vector β such that $p(X)$ is as close to 1 and 0 as possible for samples labelled with “1” and “0” respectively [45]. Mathematically, we want to maximise:

$$L(\beta) = \log\left(\prod_{i:y_i=\text{“1”}} p(x_i) \times \prod_{j:y_i=\text{“0”}} (1 - p(x_j))\right), \tag{4.4.2}$$

$L(\beta)$ is called the likelihood function [45]. This problem is the same as minimising

$$L(\beta) = -\log\left(\prod_{i:y_i=\text{“1”}} p(x_i) \times \prod_{j:y_i=\text{“0”}} (1 - p(x_j))\right), \tag{4.4.3}$$

The LASSO is derived by adding a penalty term to equation4.4.3, that is

$$L(\beta) + \lambda||\beta||_1, \tag{4.4.4}$$

where $||\beta||$ is the ℓ_1 norm defined by $||\beta|| = \sum|\beta|$, λ is the turning parameter selected by cross-validation to minimise the misclassification error. Notice that when $\lambda = 0$, we have the ordinary logistic regression, and as λ increases, more and more variables would be shrunk to zero [44].

From the LASSO model, we calculated the weighted peak area using:

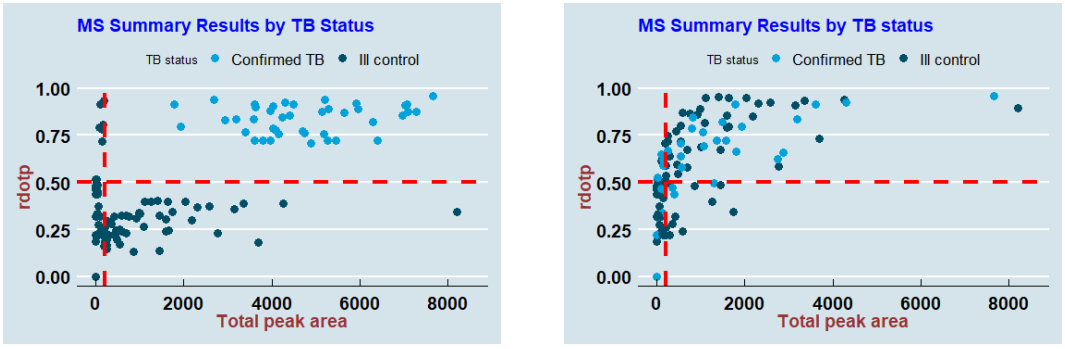
$$\text{weighted peak area} = \sum_i P_i\beta_i$$

where i is the fragment ion i, P_i is the peak area of fragment ion i derived from CFP-10 peptide, and β_i is the coefficient of LASSO corresponding to fragment ion i.

4.5 Results

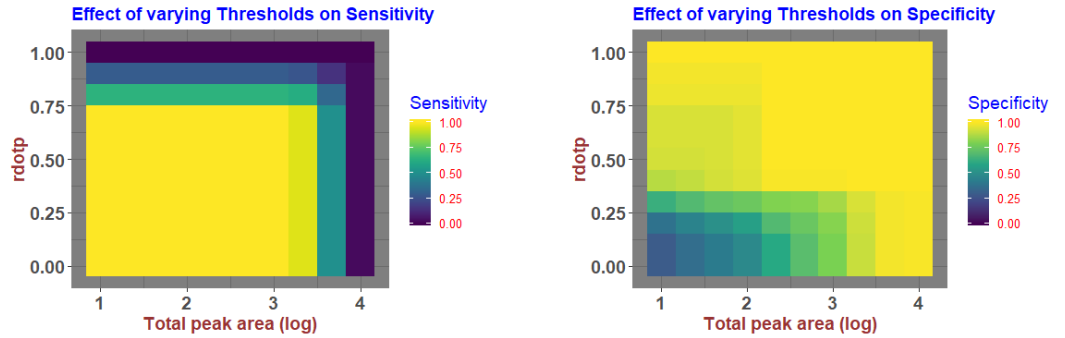
4.5.1 Two Threshold Criterion

The effect of varying threshold values on sensitivity and specificity values was explored using the two datasets. Total peak areas from 10 to 4 000 and rdotp value between 0 and 1, with step 0.1 was used.



(a) Perfect dataset (b) Imperfect dataset

Figure 4.2: Plot of total peak areas and rdotp values of each subject. Dotted red lines represent the threshold values. Proposed case definition for CFP-10 positive is total peak area > 200 and rdotp > 0.5. Notice that the two threshold lines divided each graph into four quadrants. A sample is test positive if it falls in the top right region; otherwise, it is test negative.

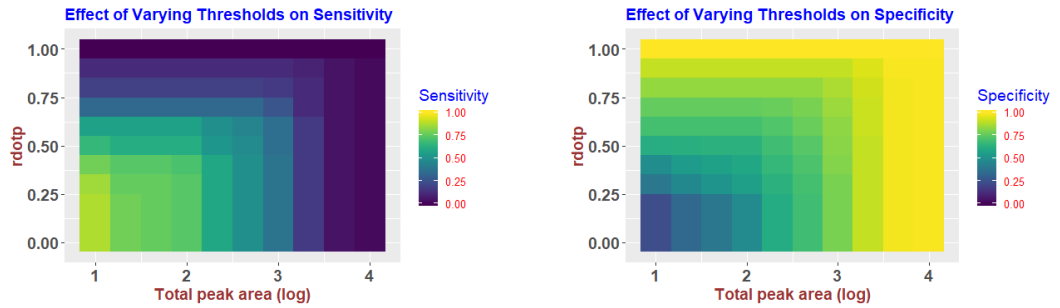


(a) Sensitivity (b) Specificity

Figure 4.3: Threshold effect on sensitivity and specificity using perfect dataset.

In Figure 4.3a, nearly all the peak areas considered yield a very high sensitivity when rdotp value is less than 0.7. However, when rdotp value is higher than 0.7, sensitivity rapidly drops. This is because in Figure 4.2a, as the vertical threshold moves from left to right up, to about 2 000, only a few subjects who are confirmed TB will move from TB region to non-TB region, but when rdotp threshold moves from 0.5 upward, many subjects would move from TB region to non-TB region, which accounts for the rapid fall in sensitivity observed in Figure 4.3a.

As the sensitivity in Figure 4.4 approaches 0, the specificity increases towards 1. When the threshold for peak area is reduced to about 100, with a reasonable rdotp threshold, adequate sensitivity and specificity are attained.

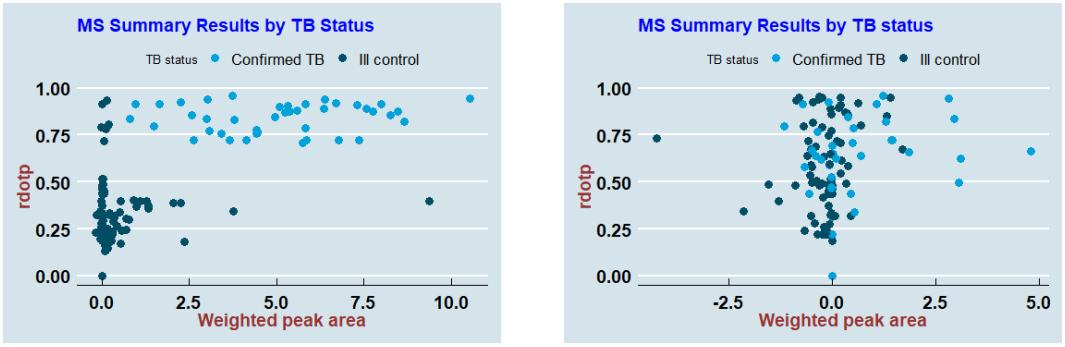


(a) Sensitivity (b) Specificity

Figure 4.4: Threshold effect on sensitivity and specificity using imperfect dataset.

4.5.2 The LASSO Method

Here, the weighted peak area is used instead of the total peak area. In Figure 4.5a, we can easily separate confirmed TB from ill control using the two-threshold criterion, and the effect of varying the thresholds are presented in Figure 4.6. With rdotp value of about 0.6 and weighted peak area between 0 and 1, we are guaranteed to have a perfect diagnostic. On the other hand, in Figure 4.7, no combination of threshold would give sensitivity and specificity both greater than 70%.

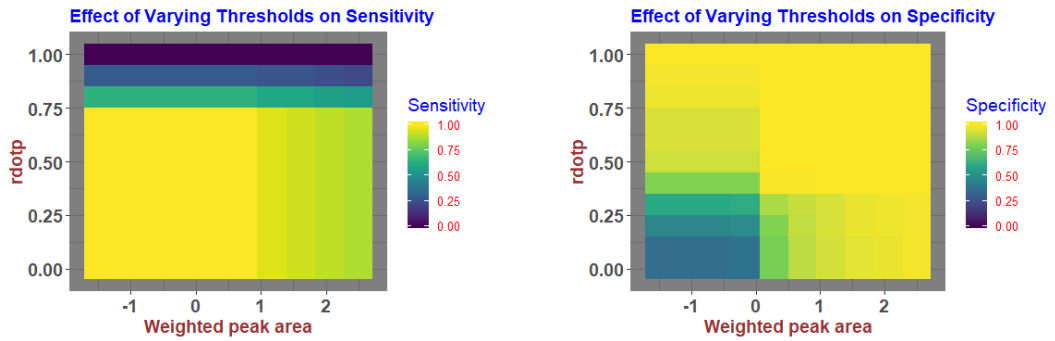


(a) Perfect dataset (b) Imperfect dataset

Figure 4.5: Plot of total peak areas and rdotp values of each subject.

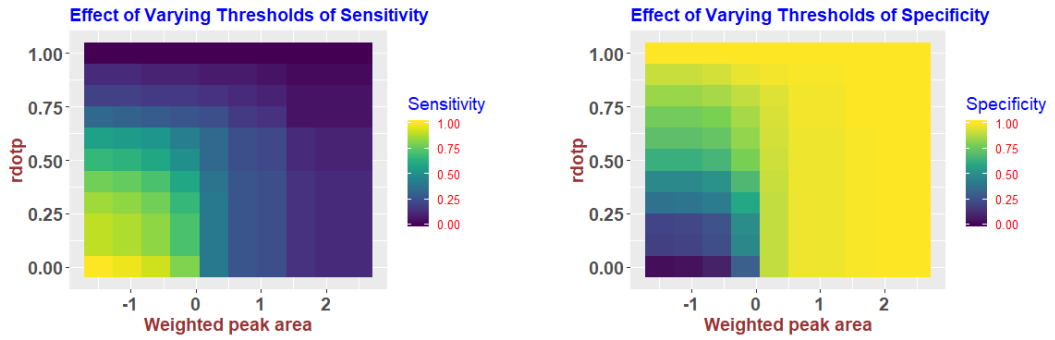
4.6 Discussion

Accurate and reliable diagnostic tests that rely on easy to collect samples are urgently needed to support TB treatment and elimination programs. Currently, available diagnostic tests perform poorly, especially in children and patients with paucibacillary TB (low quantity of TB bacteria). Diagnostic



(a) Sensitivity (b) Specificity

Figure 4.6: Threshold effect on sensitivity and specificity using perfect dataset.



(a) Sensitivity (b) Specificity

Figure 4.7: Threshold effect on sensitivity and specificity using imperfect dataset.

tests that yield quantitative results are crucial to monitor the response of patients to TB treatment.

The MRM-MS assay described here has the potential of rapidly detecting and quantifying the amount of a TB specific antigen, CFP-10, in serum samples of patients infected with TB. The assay is especially important, if effective, in testing children and sick people that cannot produce sputum We explored an ongoing project in which the idea was to assemble samples from multiple cohort studies, shipped to Tulane University, where the assay is being used on samples. Although this project is currently in progress and it is not at the stage of reporting its findings, however in this study we provided some routine analysis to explore different algorithms to convert multi-variable data into a diagnostic tool.

The summary data from MS has several variables with moderate collinear-

ity. Even though the LASSO method aims to select variables that efficiently explains the response variable in a data, our result shows that it adds little or no information to data generated by a well-understood biological process. A recent study by Van Calster et al. [46] showed that the shrinkage methods, including the LASSO model, might not improve performance in a study with small sample size or low ratio of the number of observations to the number of variables. Hence, the naive way of calculating peak area (simply summing individual peak area) is just as good.

Another stated goal of the diagnostic marker project is to investigate the utility of the MS platform in assessing treatment response – but at this point, we are not ready to consider such an evaluation because we have to understand the primary diagnostic performance better at this stage. Also, it will be a challenge that most of the patients appear to respond very favourably to treatment – as evidenced by the fact that most become culture-negative after two months. Therefore, it will be challenging finding enough patients with a poor response to be able to characterise these patients differences relative to those who respond very rapidly to treatment.

Chapter 5

Managing TB Treatment Register Data

5.1 Introduction

TB in South Africa is majorly managed within the public health care system, which is made up of the primary, secondary, and tertiary health care facilities. The primary health care facility provides diagnosis and treatment services, which are mostly provided by nurses. Diagnosis and treatment of complex cases are handled by the secondary and tertiary level facilities. In practice, standard clinical record form for all patients diagnosed with TB are completed by nurses and the summary data is then recorded into the facility paper-based registers for TB. This clinical record form was developed according to the check-list benchmark for TB surveillance and registration system provided by the WHO Global Task Force on TB Impact Management [7]. Summary data recorded in the paper-based register consist of demographic, clinical, and laboratory data. Summary data are sent from the facility level to the sub-district level, then forwarded to the district level. The district level will collate data collected from the sub-district levels and then forward it to the provincial level, before submission to the National Department of Health (NDoH), where all provincial-level data are collated and stored in the national electronic TB register [7].

For historical reasons, data about drug-sensitive and drug-resistant TB treatment is stored in structurally very different databases. The database for

drug-sensitive TB has fewer tables and shorter periods for each record, with very little follow-up details, compared to the database for drug-resistant TB (known as EDRweb). In particular, the drug-resistant database has a significantly advanced relational database design with multiple tables holding data in a more formally 'normalised' format. Normalisation is a structural way of organising data in a relational database. This process includes dividing a database into tables (entities) and then establish relationships between the tables to reduce redundancy and improve data integrity [47].

Currently, the Desmond Tutu TB Centre (DTTC) has an agreement with the NDoH to investigate key treatment outcomes and predictors of treatment outcomes, among patients being treated for DR-TB. Data available at the national database, EDRweb was transferred to the DTTC. The EDRweb is a database for DR-TB surveillance, program monitoring and evaluation. During the work on this thesis, in an ongoing collaboration between SACEMA and DTTC, SACEMA was tasked to explore the dataset and generate systematic estimates of treatment outcome frequencies. The principal activities were to:

1. Provide a report on the non-trivial structure of the EDRweb, including blocks of SQL and R code which support extracting human-readable views of data generated from linking multi-purpose lookup tables which make a direct reading of much of the structure data difficult;

2. Provide a specific data map which converts the EDRweb to a format consistent with the structure that has been adopted for the DS-TB database;

3. Provide specific descriptive analysis of the EDRweb.

5.2 Structure of Treatment Register for Drug-Resistant TB

The EDRweb has 41 tables, containing data information for 111 102 patients uniquely identified by person ID. Each patient may have multiple TB treatment episodes, uniquely identified by episode ID. In total, there are 124 557 episodes. The episode ID is the most important piece (primary key) that

links different tables together.

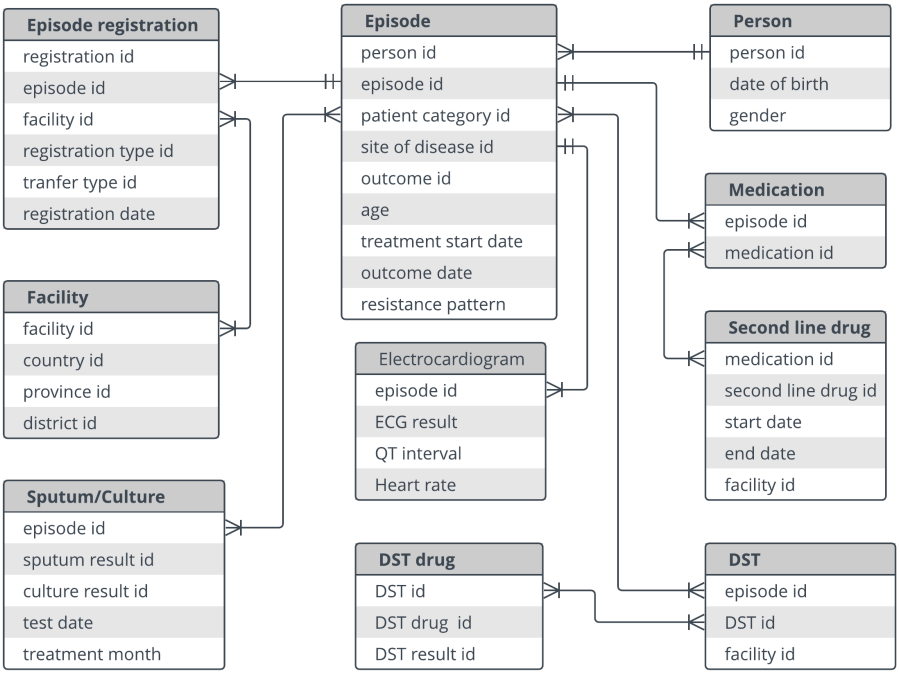


Figure 5.1: Entity related diagram of a subset of the EDR database

A proper understanding of how tables are linked together within the database is key to successful extraction of meaningful information of interest from the database. The most important tables in the database, as shown in Figure 5.1, are tables for:

- **Person (table name, tblPerson)**: This table keeps track of the core information on patients which is usually not updated over time, including date of birth and sex. If a patient is treated for more than one episode of DR-TB, then there should not be a new row in the person table – rather just a new treatment episode (see treatment episodes table below). In practice, a new row may be erroneously added to the person table, and some data cleaning would need to be performed to see whether such duplications can be detected and corrected;
- **Episode (tblEpisode)**: Each row in the episode table is intended to capture stable information about a period of treatment of a particular patient in the person table. Note that ideally, if someone is treated

more than once, there should be one row in the person table and multiple rows in the episode table;

- **Episode registration (tblEpisodeRegistration)**: This table caters to the fact that people move between facilities, and even between provincial healthcare systems, during their long course of treatment. For each facility at which a patient is being taken care of, there should be an episode registration. If someone moves from one facility to the other within a treatment episode, there should be multiple rows of episode registration associated with that episode;

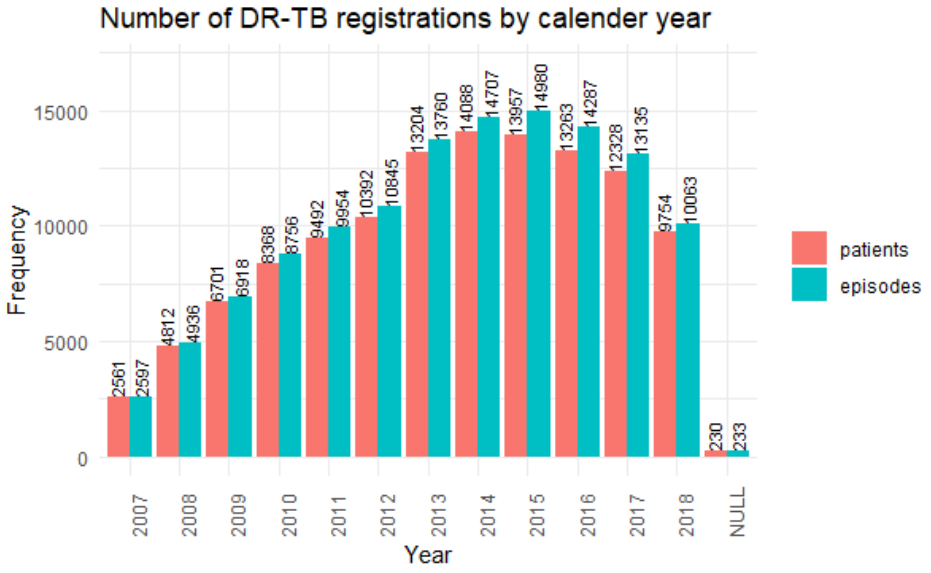


Figure 5.2: Number of patients registered in the EDRweb from 2007 to 2018 by calendar year.

- **Sputum table (tblPatientSputum)**: This table contains sputum and culture results from month 0 to month 24;

- **Drug susceptibility testing (DST) results (tblPatientDSTDrug)**: Patients may provide multiple specimens which are sent for resistance pattern testing, even during a single treatment episode. Each time a drug resistance pattern test is reported, this generates a row in this table, and this row is linked to the corresponding episode in the episode table;

- **Electrocardiogram (tblPatientECGResult)**: A significant complication of some of the DR-TB drugs is a form of cardiac arrhythmia (irregular heartbeats), and so this table contains ECG data;
- **Second-line drug (tblPatientSecondLineDrug)**: Patients reported in the DR-TB database are being treated with a different combination of second-line drugs because they have acquired resistance to the first-line drugs. This table captures the list of drugs used to treat each patient during each treatment episode; and
- **Facility (tblFacility)** This table records the location, including facility, sub-district, district, and province associated with each episode registration.

5.2.1 Lookup Tables

The fundamental purpose of a lookup table is to support data integrity by forcing data fields to have values from a constrained list of officially allowed values. For example – treatment outcomes must be either cured, treatment completed, treatment failure, died or loss to follow-up. This means that outcome cannot be recorded as 'successful', and a misspelt word that can cause confusion when treatment outcomes of a particular kind are being searched for is avoided.

In principle, one can make a separate lookup table for each instance where this kind of constraint on data is deemed to be useful, but a generic lookup table has columns with the following structure:

- ID – 'primary key' by which this data value is referred to in other tables where the lookup constraint is to be implemented;
- Short name;
- Full name; and
- Description.

This is generically known as metadata. In the DR-TB database, the data field lookups have mainly been implemented via two generic lookup tables, tblLookupItem and EnumValues. The former provides details of var-

LookupItemID	LookupGroup	Name	Description
F4429190-4EF0-423E-A5AA-5AFAFD3B2626	OUTCOME	Treatment Failure (TF)	Treatment Failure (TF)
50864FA9-EA08-4D71-907B-657A51030BBB	OUTCOME	Still on Treatment	Still on Treatment
79AF5BF3-1E7B-4BD3-9D80-69438636E7C4	OUTCOME	Treatment Completed (TC)	Treatment Completed (TC)
CD4061FD-486E-41E3-875A-80F7134AD06B	OUTCOME	Loss to Follow-Up	Loss to Follow-Up
B8A440C6-E738-4475-8767-BF31FC31F1FD	OUTCOME	Died (D)	Died (D)
3C33AC19-C62C-414F-8D76-003C5929D1C4	OUTCOME	Cured (C)	Cured (C)
9AB04649-B7CC-441C-93FE-1949F635E24C	PATIENTCATEGORY	Other	Category IV patients who do not fit any other definition
BE7D7A5B-75A9-49CF-A71F-BF4D84769DA1	PATIENTCATEGORY	TF2	Returned to treatment after re-treatment (Category II) has failed
C2C213D4-B863-4FBC-9BAE-D1E5CA1168D2	PATIENTCATEGORY	New	Never received anti-TB drugs or received anti-TB drugs for less than 1 month
E4E0B527-7B5D-4C7A-B92F-E768EE6DDF2B	PATIENTCATEGORY	Relapse	Previously treated for TB with cured or completed outcome then diagnosed with DR-TB
44D761AC-947F-489E-A1C2-F0F83A163E3D	PATIENTCATEGORY	Treatment After Loss to Follow-Up (TAL)	Returned to treatment with confirmed DR-TB after interruption of 2 months or more
3D62550C-57A2-4C5F-A514-8B24D134271C	PATIENTCATEGORY	TF1	Returned to treatment after the first treatment (Category I) has failed

Figure 5.3: A subset of the Lookup table in the EDR database. The SQL code to produce this result is available in the appendix.

ious defined data variables, such as outcome, drugs, resistance patterns, registration type, treatment history, transfer type, and test results; while the latter provides details of various adverse reactions.

5.3 Observations

Several tables in the EDRweb were underutilised as a result of incomplete records. Varied knowledge and understanding of the staff involved in data collection is a potential cause of lack of consistency in data recording. Many treatment episodes have multiple DST tests, with an enormous number of episodes without DST test results recorded, about 40 000. About 11 000 patients have more than one treatment episodes. Since EDRweb deals with DR-TB patients, multiple episodes per patient is a reflection of patients being treated several times. The sputum results table contains sputum results from baseline up to month 36. It has about 1 million observations for 99 920 distinct episodes. This implies only a few episodes, 24 654, do not have sputum results. Some treatment months recorded in the table are inaccurate. For example, several treatment months for some observations ranges between 100 and 600. Figure 5.4 shows the improvement in microbiology testing over time. The introduction of GeneXpert in 2011 for routine test-

ing has led to an increase in the proportion of patients with microbiological confirmation of TB.

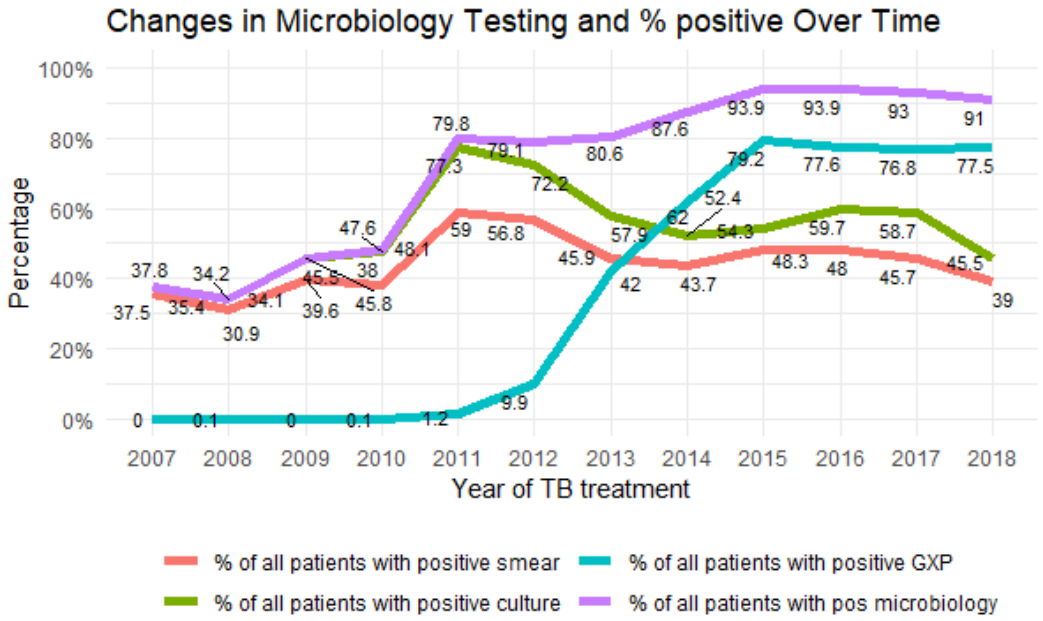


Figure 5.4: Changes in microbiology (smear/culture/Xpert) testing and proportion of positive results over time, using newly registered cohorts recovered from EDRweb. Positive microbiology is defined as a patient who is smear/culture/Xpert positive.

Out of 124 557 episodes recorded, 32 546 do not have treatment outcome recorded (about 26%). However, the quality of data before 2011 is poor, and since this dataset was extracted in 2018, those who started treatment in 2017 and 2018 are less likely to have treatment outcome recorded as they are still on treatment. Out of 96 839 episodes that initiated treatment before 2017, 13 889 have no treatment outcome recorded. Precisely 6 315 and 8 332 episodes that started treatment in 2017 and 2018 respectively, have no treatment outcome reported. In Figure 5.5, the proportion of patients with no outcome recorded continuously decreases from about 38% in 2009 to about 10% in 2015 and suddenly increases to about 19% in 2016.

Although the list of drugs at each treatment episode was well recorded for many episodes, however, the start and end dates were not fully documented. Only 46% (52 577) and 7% (7 856) have start and end dates, re-

spectively. Only about 6.4% (7 340) have both start and end dates recorded.

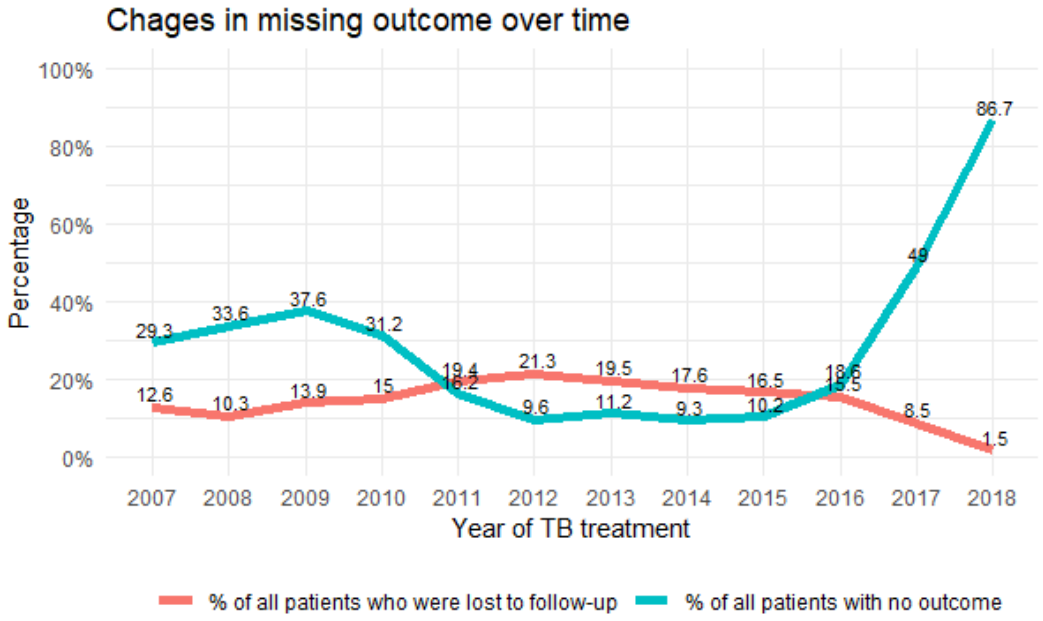


Figure 5.5: Changes in the proportion of missing outcome over time, using newly registered cohorts recovered from EDRweb.

The table containing blood results was underutilised – it contains only 32 947 observations for 8 383 unique episodes. Of those 32 947 observations, only 4 440 have CD4 counts recorded, and only 3 236 have viral load measurements. This is not a surprise because CD4 measurement in HIV patients is not required for evaluating eligibility for ART since the scaling up of ART (antiretroviral therapy) in all DR-TB patients. We cannot speculate if CD4 was not done or not recorded.

Patients on DR-TB drugs (second-line drugs) are known to experience some adverse reactions during treatment. The table to monitor adverse reactions reported contains 7 685 observations for 2 535 unique episodes; that is, only about 2% of episodes reported experienced adverse reactions. The ECG table contains 42 499 observations for only 9 752 unique episodes. However, the ECG table is newly available, it only became a requirement when bedaquiline and newer drugs were introduced.

5.4 Descriptive analysis of EDR

A comprehensive analysis of routinely collected data in EDRweb can help understand the effectiveness of DR-TB treatment and elimination programs in South Africa. We explored the EDRweb to investigate the predictors of treatment outcomes in DR-TB patients. This includes providing specific descriptive analysis of the EDRweb based on predictors including age, resistance patterns, locations, treatment durations, and regimens; and also providing estimates of treatment outcome (especially mortality) rates by various plausible predictors. Details of this analysis are provided in chapter 6.

5.5 Reshaping EDR

Previously, DTTC was tasked to work on DS-TB database to strengthen TB surveillance and monitoring in South Africa. One of the key objectives of that task was to harmonise, clean, and de-duplicate DS-TB dataset to improve its quality for management and research purposes. This task was completed at 2015, resulting in an improved DS-TB dataset containing data for over 4.7 million TB patients between 2004 and 2013. Data from 2014 to 2018 has also been added.

One of the tasks in the ongoing collaboration between SACEMA and DTTC is to reshape the EDRweb to a format consistent with the structure of the previously analysed DS-TB data. Detailed mapping of corresponding data fields is provided in the appendix A.1, which is also transferred to the data centre at DTTC, for creating a reshaping extract. A reshaped EDRweb data will facilitate the application of some routine analysis previously developed for the application to the data from DS-TB database.

5.6 Recommendations for NDoH

The EDRweb database is a well-structured database that can potentially be used to answer various vital and complex questions related to patient management and the effectiveness of TB treatment programmes, but of course, this requires accurate and consistent recording of patient's information dur-

ing clinical care. Some data appears to be more often missing than recorded, so it may be worth investing some effort in increasing user's awareness of the database capabilities. Understanding the various pressures and distractions within the TB treatment system may shed light on the reasons that useful details are often not reported. As the DR-TB database schema is quite rich and flexible, there is in principle no reason why all TB treatment cannot be recorded in a single database.

Chapter 6

Analysis of DR-TB Treatment Outcomes

6.1 Introduction

In South Africa, approximately 10 000 patients are officially treated for DR-TB every year, accounting for about 70% of all national DR-TB cases [1, 48]. Treatment of DR-TB is challenging, with treatment success of about 50%. Until recently, there has not been a coherent analysis of DR-TB treatment and control programs in South Africa. The use of the electronic drug-resistant TB register (EDRweb) for routine monitoring of DR-TB patients has provided a good data source for in-depth analysis of DR-TB treatment control programs, including measuring treatment outcomes.

6.2 Methods

Study Design

A retrospective cohort analysis of all patients treated for DR-TB that are routinely reported in EDRweb from 2007 to 2018, across all provinces of South Africa was conducted.

Data Source and Extraction

This study makes use of data from the de-identified version of EDR database received from the NDoH. The DR-TB database contains patient-level TB

treatment and follow-up records from 2007 to 2018. All data extraction was done using SQL. All SQL and R codes used in this study is available in a private GitHub repository; access will be provided upon request. Main non-trivial data extraction components of this study include:

- **Age**: EDRweb captures patients' exact date of birth, but for analytical purposes, we mainly aggregate patients into age bins defined as 0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, and 65+;
- **Patient category**: This field describes previous treatment experience of patients, see Table 6.1.

Table 6.1: Definition of TB patient categories.

Patient category	Definition
New	patients who never received anti-TB drugs or received anti-TB drugs for less than 1 month
TF1	patients who returned to treatment after the first TB treatment has failed
TF2	patients who returned to treatment after TB re-treatment has failed
Loss to follow-up	patients who returned to treatment with confirmed DR-TB after interruption of 2 months or more
Relapse	patients who were previously treated for TB with cured or completed outcome, then diagnosed with DR-TB
Other	patients who were previously treated but the outcome was not known
Unknown	patients who do not fit any other definition

- **Registration type**: This field captures whether a facility is the first to handle a given treatment episode ('new' registration), whether it is taking over treatment from another facility in the same province ('moved in') or from another province ('transfer in');
- **Site of disease**: The location of infection within the patient may vary in complex ways, but is summarised here simply as Pulmonary TB (PTB – affecting the lungs), or extra-pulmonary TB (EPTB – affecting other parts of the body);

- **Treatment outcome**: The EDRweb reported DR-TB treatment outcome as cured, treatment completed, died, treatment failure, still on treatment, loss to follow-up and unknown, defined in Table 2.2. Patients who were cured or completed treatment were classified as treatment success, those whose treatment outcome was unknown or loss to follow-up were aggregated as missing outcome. Patients whose treatment outcome was reported as still on treatment was allowed for those who started treatment in 2017 and after;

- **Drug resistance patterns**: The EDRweb uses three resistance pattern classifications (Monopoly, MDR, XDR), and this is specified for most patients. We decided to calculate a more detailed set of categories (see Table 6.2), however, due to missing data from DST, we were unable to ascertain this refined category for nearly 20% of patients;

Table 6.2: Definition and categories of DR-TB based on resistance pattern.

Resistance pattern	Definition
RifMono	resistant to rifampicin only and specifically sensitive to isoniazid
MonoPoly	resistant to one or more of isoniazid, ethambutol, pyrazinamide
MDR	resistant to rifampicin and isoniazid or resistant to rifampicin and isoniazid not specified
PreXDR	resistant to rifampicin and isoniazid and [resistant to injectable (kanamycin or amikacin or capreomycin) or resistant to a quinolone (moxifloxacin or ofloxacin or levofloxacin or gatifloxacin or ciprofloxacin)]
XDR	resistant to rifampicin and Isoniazid and [resistant to injectable (kanamycin or amikacin or capreomycin) and resistant to a quinolone (moxifloxacin or ofloxacin or levofloxacin or gatifloxacin or ciprofloxacin)]

RifMono, rifampicin-resistant TB; MonoPoly, mono or poly resistant TB; MDR, multi drug-resistant TB; XDR, extensively drug-resistant TB.

- **Drugs used**: The second-line drugs used to treat patients was recorded as 'list of drugs' in the data extract. This variable was used to record

the use of drugs in group A, B, and C as per Figure 6.1 recorded as yes or no.

Groups & steps	Medicine	
Group A: Include all three medicines	levofloxacin *OR* moxifloxacin	Lfx Mfx
	bedaquiline[2,3]	Bdq
	linezolid[4]	Lzd
Group B: Add one or both medicines	clofazimine	Cfz
	cycloserine *OR* terizidone	Cs Trd
Group C: Add to complete the regimen and when medicines from Groups A and B cannot be used	ethambutol	E
	delamanid[3,5]	Dlm
	pyrazinamide[6]	Z
	imipenem–cilastatin *OR* meropenem[7]	Ipm–Cln Mpm
	amikacin (*OR* streptomycin)[8]	Am (S)
	ethionamide *OR* prothionamide[9]	Eto Pto
	p-aminosalicylic acid[9]	PAS

Figure 6.1: Grouping of medicines recommended by WHO for more extended MDR-TB regimens. Adapted from World Health Organization [49, p. 24].

Analysis

All analyses were conducted using the R statistical program. Basic descriptive statistics, including frequency, proportions and rates, were used to explore the patient's demographic, clinical, diagnostic, treatment, and outcome characteristics.

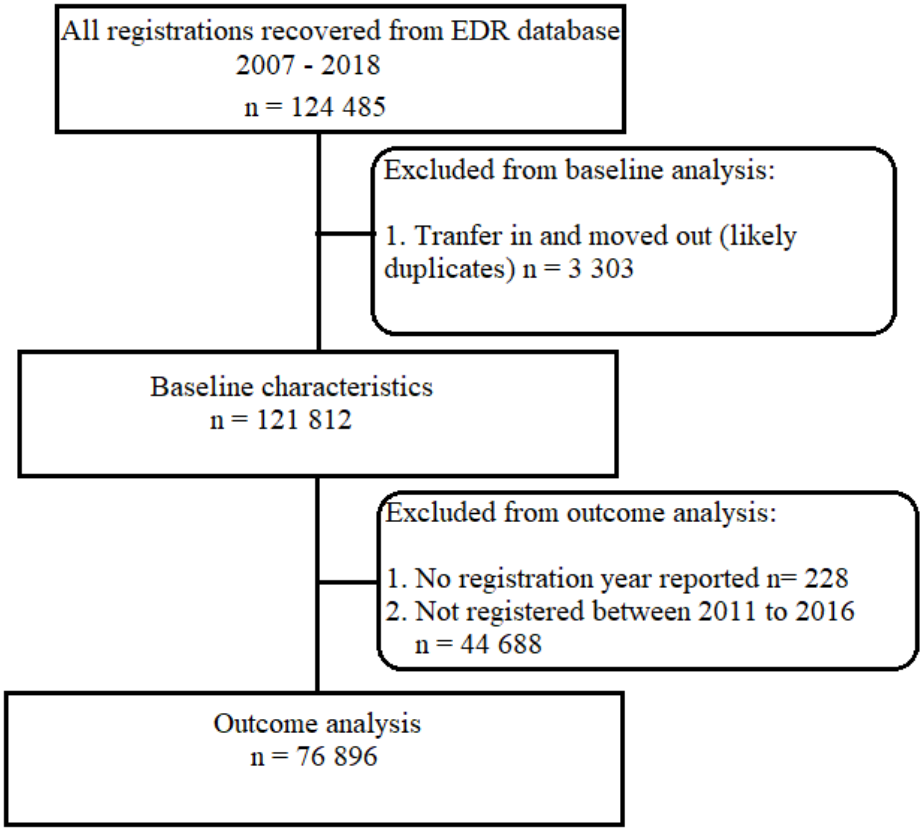


Figure 6.2: Patients included in the baseline and outcome analysis

6.3 Results

6.3.1 Baseline Demographic and Clinical Characteristics

Baseline characteristics of newly registered DR-TB patients are summarised in Table 6.3 using frequencies and proportions. One hundred and twenty-one thousand, eight hundred and twelve newly registered cohorts were recovered from the EDRweb and satisfy the inclusion criteria for baseline analysis. Approximately 55% ($n = 66\,638$) was male. The median age of all patients was 35 years (IQR: 38-44). DR-TB is most common in patients between 25-34 years (31.6%) and least common in young children below five years (1.1%). KwaZulu-Natal has most cases of DR-TB reported, at 29.6% ($n = 36\,045$) while Limpopo province has the least number of cases, at 3.3% ($n = 4\,047$).

More than half of the patients were HIV positive, at 57.8% ($n = 70\,462$) and 25% ($n = 30\,511$) HIV negatives. Of those who were HIV positive, 87.7%

Table 6.3: Baseline demographic and clinical characteristics of drug-resistant TB patients reported in EDRweb from 2007 to 2018.

Characteristic	Overall (N=121 812)	Proportion
Gender		
Male	66 638	54.7%
Female	54 932	45.1%
Unknown	242	0.2%
Age bin		
0-4	1 379	1.1%
5-14	2 374	1.9%
15-24	15 601	12.8%
25-34	38 474	31.6%
35-44	35 192	28.9%
45-54	18 938	15.5%
55-64	7 075	5.8%
65+	2 617	2.1%
Unknown	161	0.1%
Province		
Eastern Cape	22 097	18.1%
Free State	5 686	4.7%
Gauteng	13 553	11.1%
KwaZulu-Natal	36 045	29.6%
Limpopo	4 047	3.3%
Mpumalanga	8 934	7.3%
North West	5 643	4.6%
Northern Cape	4 740	3.9%
Western Cape	21 067	17.3%
HIV status		
Positive	70 462	57.8%
On ART	61 772	87.7%
Not on ART	3 529	5.0%
Unknown	5 161	7.3%
Negative	30 511	25.0%
Unknown	20 839	17.1%

Characteristic	Overall (N=121 812)	Proportion
Treatment Year		
2007	2 571	2.1%
2008	4 888	4.0%
2009	6 856	5.6%
2010	8 669	7.1%
2011	9 812	8.1%
2012	10 734	8.8%
2013	13 582	11.2%
2014	14 500	11.9%
2015	14 529	11.9%
2016	13 739	11.3%
2017	12 179	10.0%
2018	9 525	7.8%
Unknown	228	0.2%
Site of disease		
PTB	111 312	91.4%
EPTB	1 414	1.2%
Unknown	9 086	7.5%
Patient category		
New	51 419	42.2%
Re-treatment	65 782	54.0%
Unknown	4 611	3.8%
Resistance type		
Mono or poly	3 294	2.7%
MDR	106 396	87.3%
XDR	7 301	6.0%
Unknown	4 821	4.0%
Pre-treatment microbiology		
Positive	96 516	79.2%
Negative	4 247	3.5%
Unknown	21 049	17.3%

were on ART and 5% were not on ART. About 17% ($n = 20\,839$) have unknown HIV status. Figure 6.3 shows the changes in HIV testing and ART coverage over time. HIV testing for all registered DR-TB cases rapidly improved over time, with only about 3% having unknown HIV status towards 2018. Proportion of all HIV infected DR-TB patients who were initiated on ART also rapidly increases from about 21% up to about 90% in 2015. The use of EDRweb for routine surveillance of DR-TB treatment programs was implemented in 2013. The year 2007 to 2009 was considered as pilot while 2010 and 2011 was rollout phase.

Pulmonary TB is the most common diagnosis reported, approximately 91% ($n = 111\,312$), with only 1.2% ($n = 1\,414$) reported as having exclusive extra-pulmonary TB. Note that many patients with extra-pulmonary TB also suffer from pulmonary TB, and therefore not classified as extra-pulmonary TB. Just over half of all patients, 54% were reported to have been previ-

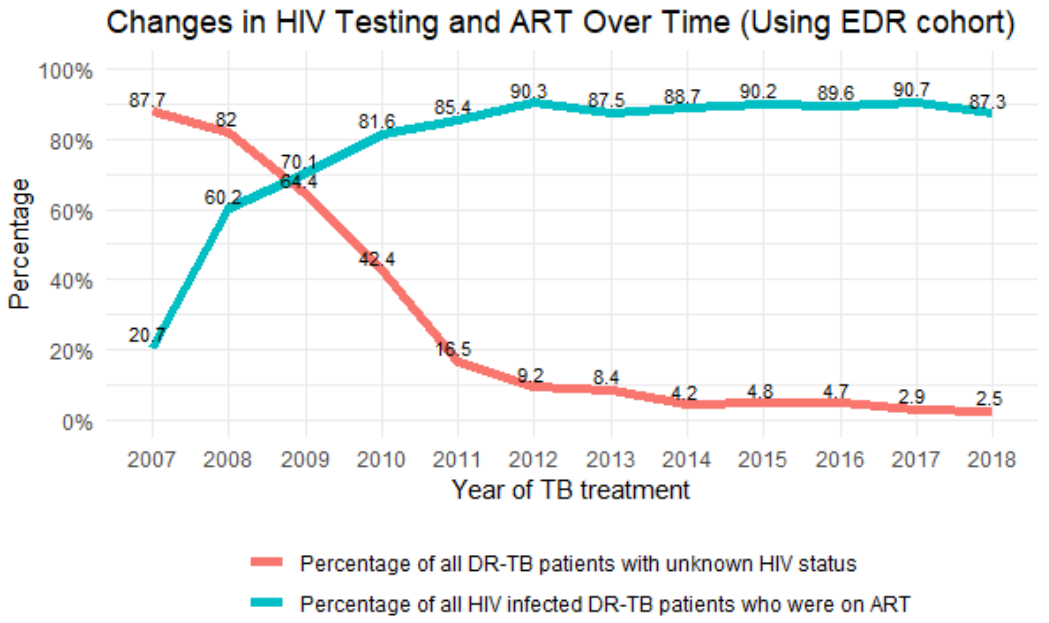


Figure 6.3: Trend of HIV control programs in South Africa from 2007 to 2018. ART, antiretroviral therapy.

ously treated for TB, and 42.2% were categorised as new cases of DR-TB. About 87% of cases in the cohort were MDR, 2.7% were mono or poly resistant, and 6.0% were XDR. Positive pre-treatment microbiology (smear/-culture/Xpert) was found in 79.2% ($n = 96\,516$), negative results was estimated as 3.5% ($n = 4\,247$), while the rest were either contaminated or unreported.

6.3.2 Case Notification Rates of DR-TB

Case notification rate (CNR) of DR-TB per 100 000 population-time by gender, age bins, and province from 2011 to 2016 are presented in Table 6.4. The CNR is higher in male than in females over time. The age group 35-44 have highest CNR (from 35 in 2011 to 41 in 2016), while children have the lowest CNR. The Limpopo province has low CNR over time while the Eastern Cape, Western Cape, Northern Cape and KwaZulu-Natal have very high CNR of DR-TB. Almost all categories have significant increase in CNR from 2012 to 2013.

Table 6.4: Estimated case notification rates of drug-resistant TB per 100 000 population over time in South Africa.

	Year of TB Treatment					
	2011	2012	2013	2014	2015	2016
Gender						
Male	20	22	28	31	31	29
Female	18	19	23	24	23	21
Age bin						
0-4	2	3	3	3	3	3
5-14	3	3	3	3	3	3
15-24	13	15	18	19	18	17
25-34	35	38	45	47	44	41
35-44	45	46	59	61	61	54
45-54	31	33	43	44	44	42
55-64	16	16	22	23	24	24
65+	6	6	10	11	13	12
Province						
Eastern Cape	26	23	39	40	42	40
Free State	17	28	26	25	21	20
Gauteng	7	5	10	13	15	13
KwaZulu-Natal	31	37	41	41	37	33
Limpopo	6	6	9	11	10	9
Mpumalanga	19	16	30	28	28	25
North West	11	13	16	21	22	21
Northern Cape	41	35	34	35	49	42
Western Cape	33	39	36	36	34	35

6.3.3 Treatment Outcomes

Treatment outcome recorded for newly registered cohorts are summarised in Table 6.5. Out of the 76 896 patients treated for DR-TB between 2011 to 2016, 30.9% (18.1% loss to follow-up and 12.7% not evaluated) have missing outcome. Treatment success was recorded in 44% (cured, 33.7% and treatment completed, 10.7%), 20.5% died during treatment, and 4.3% failed treatment. Proportion of treatment success increases from 39% in 2011 to

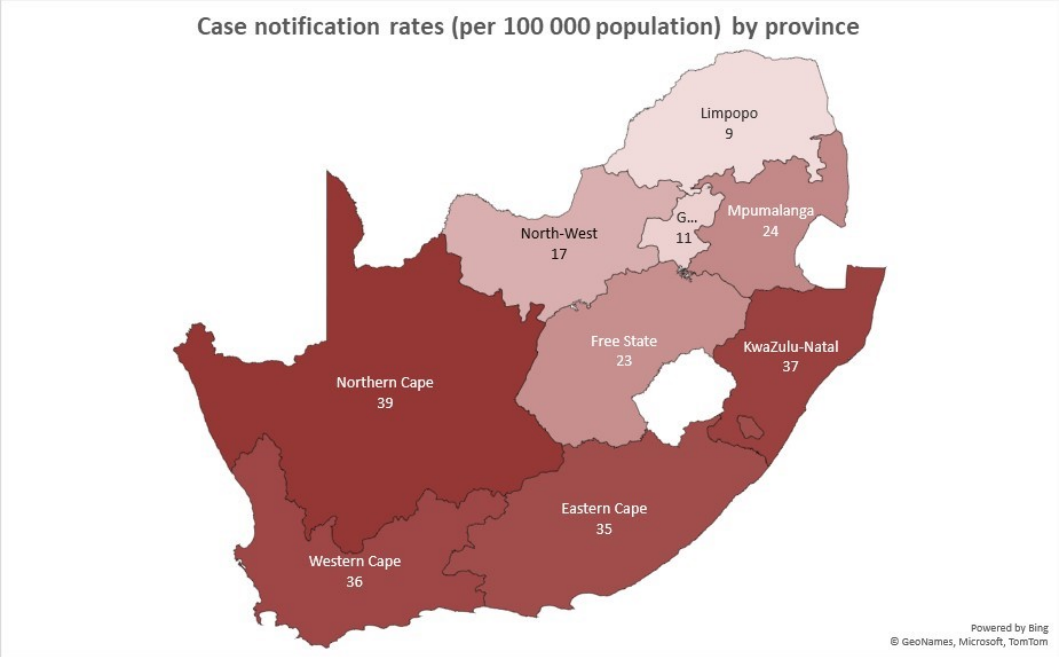


Figure 6.4: Case notification rates of DR-TB by province

47.5% in 2015. The significant decrease in counts observed in 2016 may reflect incomplete data.

Table 6.5: Treatment outcome over time

Outcome	Year of TB treatment						Total
	2011 (N=9 812)	2012 (N=10 734)	2013 (N=13 582)	2014 (N=14 500)	2015 (N=14 529)	2016 (N=13 739)	(N=76 896)
Success	3 828 (39.0%)	4 680 (43.6%)	5 852 (43.1%)	6 851 (47.2%)	6 907 (47.5%)	6 004 (43.7%)	34 122 (44.4%)
Cured	2 715 (27.7%)	3 393 (31.6%)	4 101 (30.2%)	5 244 (36.2%)	5 556 (38.2%)	4 892 (35.6%)	25 901 (33.7%)
Completed	1 113 (11.3%)	1 287 (12.0%)	1 751 (12.9%)	1 607 (11.1%)	1 351 (9.3%)	1 112 (8.1%)	8 221 (10.7%)
Died	1 976 (20.1%)	2 178 (20.3%)	2 898 (21.3%)	3 078 (21.2%)	3 072 (21.1%)	2 528 (18.4%)	15 730 (20.5%)
Failed	481 (4.9%)	521 (4.9%)	615 (4.5%)	638 (4.4%)	636 (4.4%)	415 (3.0%)	3 306 (4.3%)
Missing	3 527 (35.9%)	3 355 (31.3%)	4 217 (31.0%)	3 933 (27.1%)	3 914 (26.9%)	4 792 (34.9%)	23 738 (30.9%)
LTFU	1 907 (19.4%)	2 291 (21.3%)	2 645 (19.5%)	2 554 (17.6%)	2 397 (16.5%)	2 124 (15.5%)	13 918 (18.1%)
Not evaluated	1 620 (16.5%)	1 064 (9.9%)	1 572 (11.6%)	1 379 (9.5%)	1 517 (10.4%)	2 668 (19.4%)	9 820 (12.7%)

Treatment Outcome by Demographic and Baseline Predictors

The distribution of treatment outcome by various plausible predictors is presented in Table 6.6. Treatment success in male was 42.6%, with 20.2% deaths. In females, treatment success was estimated as 46.5%, with 20.8% deaths. Overall, the proportion of children (age 0-4 and 5-14) reported as treatment success was approximately 66% and 65%, respectively; with the lowest treatment success rate recorded in patients older than 64 years, at

32.5%. The proportion of death in very young children was 3.2%; this increases steadily with age and year of TB treatment up to 39.2% in patients older than 64 years (see Figure 6.5). Treatment failure recorded in young children was very low, 0.9%.

Table 6.6: Treatment outcome reported in EDR from 2011 to 2016 by demographic and baseline predictors. Each cell represents the proportion of treatment outcome in each row, with a binomial confidence interval.

Predictor	Success (N=34122)	Died (N=15730)	Failed (N=3306)	Missing† (N=23738)
Sex				
Male	0.426 (0.421, 0.430)	0.202 (0.198, 0.206)	0.043 (0.041, 0.045)	0.329 (0.325, 0.334)
Female	0.465 (0.460, 0.471)	0.208 (0.204, 0.212)	0.043 (0.040, 0.045)	0.284 (0.280, 0.289)
Difference	-0.039 (-0.046, 0.032)	-0.006 (-0.002, 0.004)	0.000 (-0.002, 0.004)	0.045 (0.038, 0.052)
Age bin				
0-4	0.656 (0.622, 0.689)	0.032 (0.021, 0.046)	0.009 (0.003, 0.018)	0.303 (0.272, 0.336)
5-14	0.635 (0.609, 0.660)	0.097 (0.082, 0.114)	0.029 (0.021, 0.039)	0.239 (0.217, 0.262)
15-24	0.449 (0.439, 0.459)	0.144 (0.137, 0.151)	0.052 (0.048, 0.057)	0.355 (0.346, 0.365)
25-34	0.420 (0.414, 0.426)	0.201 (0.196, 0.206)	0.044 (0.042, 0.047)	0.335 (0.329, 0.341)
35-44	0.444 (0.437, 0.450)	0.211 (0.206, 0.216)	0.043 (0.041, 0.046)	0.302 (0.296, 0.308)
45-54	0.468 (0.459, 0.477)	0.223 (0.215, 0.230)	0.044 (0.040, 0.048)	0.265 (0.257, 0.273)
55-64	0.438 (0.424, 0.453)	0.280 (0.267, 0.294)	0.034 (0.029, 0.040)	0.248 (0.235, 0.261)
65+	0.325 (0.302, 0.348)	0.392 (0.368, 0.416)	0.016 (0.011, 0.024)	0.267 (0.246, 0.289)
Province				
Eastern Cape	0.354 (0.346, 0.362)	0.287 (0.280, 0.295)	0.081 (0.077, 0.086)	0.277 (0.270, 0.285)
Free State	0.389 (0.374, 0.405)	0.222 (0.209, 0.236)	0.022 (0.017, 0.027)	0.367 (0.352, 0.383)
Gauteng	0.416 (0.405, 0.427)	0.187 (0.179, 0.196)	0.014 (0.011, 0.016)	0.383 (0.372, 0.394)
KwaZulu-Natal	0.539 (0.533, 0.546)	0.172 (0.167, 0.177)	0.033 (0.030, 0.035)	0.256 (0.250, 0.262)
Limpopo	0.507 (0.488, 0.526)	0.142 (0.130, 0.156)	0.054 (0.046, 0.063)	0.297 (0.280, 0.314)
Mpumalanga	0.452 (0.439, 0.464)	0.227 (0.217, 0.238)	0.045 (0.040, 0.050)	0.276 (0.265, 0.287)
North West	0.500 (0.484, 0.516)	0.237 (0.223, 0.251)	0.025 (0.020, 0.031)	0.238 (0.224, 0.252)
Northern Cape	0.353 (0.335, 0.371)	0.275 (0.258, 0.292)	0.076 (0.066, 0.086)	0.297 (0.280, 0.314)
Western Cape	0.388 (0.380, 0.397)	0.158 (0.152, 0.164)	0.040 (0.036, 0.043)	0.414 (0.406, 0.423)
HIV status				
Positive	0.449 (0.444, 0.453)	0.236 (0.233, 0.240)	0.041 (0.040, 0.043)	0.274 (0.270, 0.277)
Negative	0.497 (0.490, 0.504)	0.138 (0.133, 0.143)	0.051 (0.049, 0.055)	0.314 (0.307, 0.320)
Difference	-0.048 (-0.056, -0.040)	0.098 (0.092, 0.104)	-0.010 (-0.013, -0.006)	-0.040 (-0.048, -0.033)
Site of disease				
PTB	0.455 (0.451, 0.458)	0.206 (0.203, 0.209)	0.044 (0.043, 0.046)	0.295 (0.291, 0.298)
EPTB	0.506 (0.472, 0.539)	0.190 (0.164, 0.217)	0.014 (0.007, 0.024)	0.291 (0.261, 0.322)
Difference	-0.051 (-0.085, -0.0171)	0.017 (0.206, 0.190)	0.030 (0.022, 0.039)	0.004 (-0.027, 0.035)
Treatment history				
New	0.481 (0.475, 0.486)	0.184 (0.180, 0.188)	0.036 (0.034, 0.038)	0.300 (0.295, 0.305)
Re-treatment	0.431 (0.426, 0.436)	0.229 (0.225, 0.233)	0.051 (0.049, 0.053)	0.289 (0.285, 0.294)
Difference	0.050 (0.043, 0.057)	-0.211 (-0.215, -0.207)	-0.015 (-0.018, -0.012)	0.011 (0.004, 0.017)
Resistance type				
Mono or poly	0.502 (0.481, 0.522)	0.096 (0.084, 0.109)	0.031 (0.024, 0.039)	0.372 (0.352, 0.392)
MDR	0.462 (0.459, 0.466)	0.200 (0.197, 0.203)	0.036 (0.034, 0.037)	0.302 (0.298, 0.305)
XDR	0.275 (0.262, 0.288)	0.350 (0.336, 0.364)	0.168 (0.157, 0.179)	0.208 (0.196, 0.220)

KwaZulu-Natal has the highest proportion of treatment success (53.9%), followed by Limpopo and North-West province, at 50.7% and 50%, respectively. The Eastern Cape province is associated with high proportion of deaths and treatment failure, at 28.9% and 8.1%, repectively, followed by

Northern Cape, 27.5% deaths and 7.6% treatment failure. The proportion of missing outcome in the Western Cape province is about 41% of the patients treated in the province. HIV negative patients have low proportion of deaths, at 13.8% compared to HIV positive patients where the proportion of deaths was 23.6%. The proportion of treatment success was about 5% lower in HIV positive patients compared to the HIV negative patients. Treatment success in patients diagnosed with exclusive EPTB was about 51%, with 19% deaths while in PTB patients, treatment success was about 46% and 24% deaths. Approximately 48% of patients without history of TB treatment were successfully treated and 18.4% died.

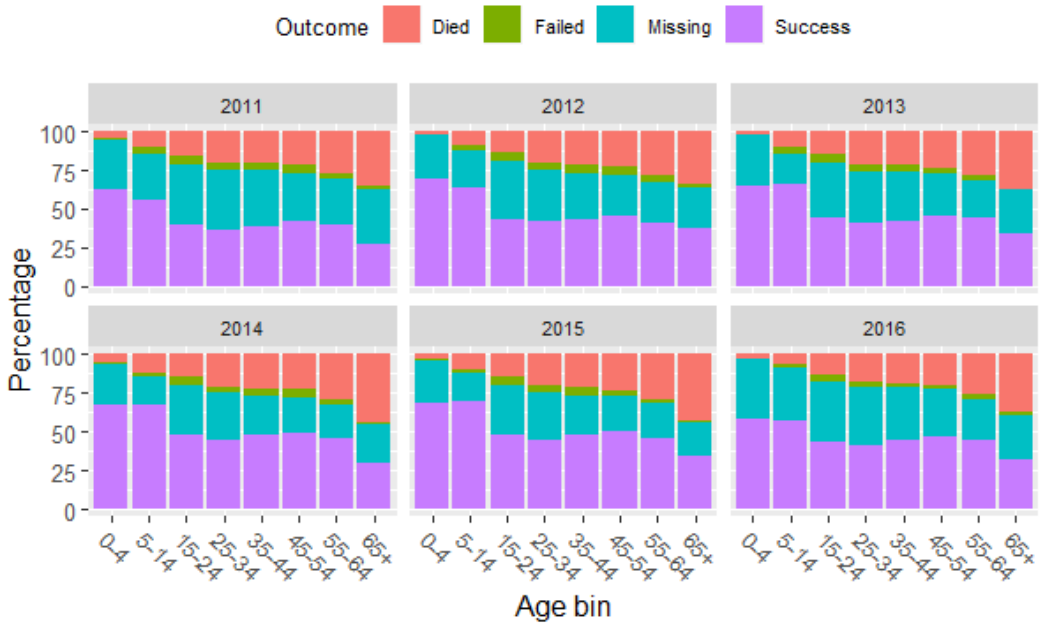


Figure 6.5: Treatment outcome of all patients treated for DR-TB in each year between 2011 and 2016 as a function of age distribution.

Treatment success in patients diagnosed with mono or poly resistant TB was 50.2%, 9.6% died, and 3.1% failed treatment. A significant proportion, 35% of patients diagnosed with XDR TB died during treatment, with treatment success in only 27.5%. Out of those diagnosed with MDR TB, 46.2% were successfully treated, 20% died, and 3.6% failed treatment. Figure 6.6 shows the significant improvement over time in the treatment of patients

diagnosed with XDR TB. Over time, death rate in XDR patients steadily decreases, with steady increase in the treatment success rate.

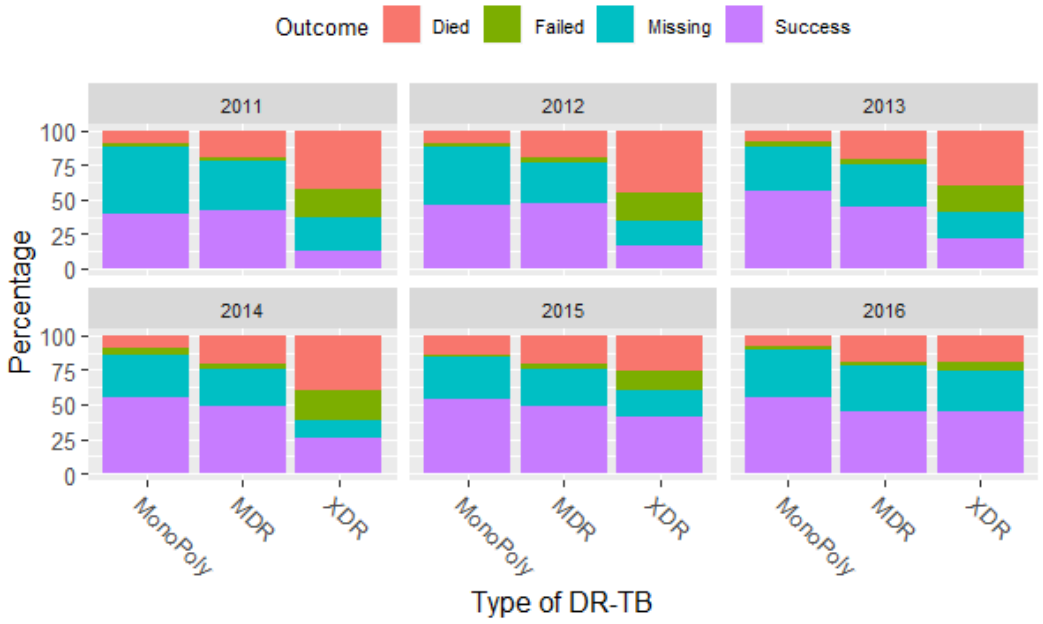


Figure 6.6: Treatment outcome of all patients treated for DR-TB in each year between 2011 and 2016 as a function of resistance pattern.

6.4 Discussion

Since 2007, South Africa has adopted the use of the EDRweb for DR-TB surveillance, program monitoring and evaluation. In this study, we extracted data for all patients routinely treated for DR-TB in South Africa and reported in EDR between 2007 and 2018. Baseline analysis includes newly registered cases from 2007 to 2018 while outcome analysis includes only newly registered cases from 2011 to 2016. A total of 121 812 newly registered cases of DR-TB was notified in EDRweb between 2007 and 2018.
Patients whose treatment outcome was missing or loss to follow-up was about 31%. Overall treatment success was estimated as 44.4% with 20.5% deaths. Despite having low rate treatment success, our results show that this rate has been constantly improving since 2011. One of the End TB Strategy recommended by WHO is to attain a treatment success of at least 90% out of all notified TB, comprising the drug-sensitive and drug-resistant TB,

in each country by 2025. To reach this target, continuous progression must be maintained across all provinces in South Africa to better improve the treatment success rate that we currently estimated.

Treatment success and mortality are associated with HIV status, history of TB treatment, age and resistance pattern. In HIV positive patients, treatment success was 44.9% with 23.6% mortality, and for HIV negative patients, treatment success was 49.7% with 13.8% mortality. Children have a higher treatment success rate and lower mortality compared to adults who are associated with low treatment success rate and a very high death rate. The resistance pattern XDR is associated with: very low treatment success rate, about 27.5%; high mortality rate, about 35%; and a high rate of treatment failure, at 16.8%. A study to further investigate each predictor is currently in progress. The ongoing study will make full use of the data that we have extracted and harmonised.

This study has some limitations: (1) The significant proportion of patients with a missing outcome is the greatest limitation of this study and (2) The EDR contains data collected from patients routinely treated for DR-TB. People living with DR-TB and not notified are not reported in the EDR hence they are not included in the dataset. Those who died before registration were also excluded.

To conclude, the diagnosis with XDR and old age are the strongest indicators of mortality during treatment and poor rates of treatment success. Predictors identified in this study should be utilised to improve patient management.

Chapter 7

Conclusion

The aim of this investigation was to better understand the spectrum of tuberculosis (TB) disease in ways that facilitate the investigation of diagnostic and prognostic indicators. More specifically, the study objectives included understanding the concepts in TB diagnosis and treatment, designing a study for biomarker evaluation, analysing the data from a biomarker evaluation study, exploring TB treatment data structures, and analysing drug-resistant TB (DR-TB) treatment outcomes.

Diagnostic performance is often summarised by the metrics of sensitivity and specificity. However, these metrics are not robust as they vary with disease prevalence and the distribution of times since infection. Other metrics, including predictive values and likelihood ratios, are also useful and should be considered. A good understanding of each of these metrics, in addition to a careful consideration of the context in which the test is to be used, can inform better clinical decisions.

When designing a study of the performance of a potential new diagnostic marker, one would usually set out to compare the new test to a gold standard. However, the lack of a perfect gold standard TB diagnostic, to obtain the true disease status of subjects, is a major source of uncertainty. There is an ongoing collaborative study between Desmond Tutu TB Centre (DTTC) and a US based collaborator aiming to evaluate a potential new marker of TB disease and treatment response. While this project's results are not ready for publication, we analysed hypothetical (simulated) data with the precise structure of the data being generated within this study. Given

multi-dimensional data driven by a well-understood biological process, a human-driven search for evidence of a diagnostic within the data is seen to be on par with a more 'sophisticated' automated search. In other words, a simple analysis is often more informative or akin to a complex method for understanding simple data.

In the South African National Department of Health TB treatment programme, data structures for recording treatment of drug-sensitive TB (DS-TB) are quite different, and separate, from data structures for recording treatment of DR-TB. We created a convenient mapping tool that allows people to query the DR-TB data efficiently. This will facilitate a direct application of any previously developed analysis for the DS-TB data to the DR-TB data by simply reshaping the data into a compatible format. TB treatment is very challenging, as patients are subjected to a long and unpleasant course of treatment. Although a significant number of the DR-TB patients do not have a recorded treatment outcome, unfavourable treatment outcomes, such as death, are found to be alarmingly frequent, and to be significantly associated with HIV status, history of previous TB treatment, age, and resistance patterns. A diagnosis of 'extensively drug-resistant TB' (XDR-TB) and old age were the strongest predictors of mortality.

In various aspects of the battle against TB, it is useful to understand the fine points associated with identifying, describing, and responding to various levels of TB disease. It is clear that there needs to be much more research into the 'spectrum of TB disease', and this work has highlighted numerous conceptual and analytical subtleties which are relevant to the performance of this research. We hope that the analyses and discussions in this thesis will support the optimisation of future research design and execution.

Appendices

Appendix A

SQL and R Codes

A.1 SQL Code to Extract Data

The main SQL queries for data extraction and the R code for data manipulation are available on a private GitHub repository, https://github.com/SACEMA/EDR_Project_SACEMA_DTTC, access will be granted upon request.

A.2 Basic Analysis Code in R

The `tableby()` function, available within the `arsenal` R package was created by Heinzen et al. [50] in 2019 to produce descriptive summary statistics tables for any set of variables (whether categorical or continuous), and sometimes stratified by groups. For example, code A.1 was used to generate a subset of the descriptive statistics table presented in Table 6.3. Note that the output of code A.1 was further modified using `xtable()` function, available within the `xtable` package [51] to produce the LaTeX output in Table 6.3.

```
1 library(arsenal)
2 table_one <- tableby(~ Sex + AgeBin  + Province,
3 data = predictors_desc_stat)
4 summary(table_one, title = "predictors_desc_stat",
5 text = TRUE, test = FALSE)
```

Code A.1: Code to derive summary statistics table for sex, ag bin, and province.

	Overall (N=124484)
Sex	
- Male	68108 (54.7%)
- Female	56132 (45.1%)
- NULL	244 (0.2%)
AgeBin	
- 0to4	1403 (1.1%)
- 5to14	2407 (1.9%)
- 15to24	15987 (12.8%)
- 25to34	39319 (31.6%)
- 35to44	35944 (28.9%)
- 45to54	19348 (15.5%)
- 55to64	7253 (5.8%)
- 65+	2667 (2.1%)
- NULL	156 (0.1%)
Province	
- EASTERN CAPE	22708 (18.2%)
- FREE STATE	5793 (4.7%)
- GAUTENG	13744 (11.0%)
- KWAZULU-NATAL	36934 (29.7%)
- LIMPOPO	4164 (3.3%)
- MPUMALANGA	9246 (7.4%)
- NORTH WEST	5685 (4.6%)
- NORTHERN CAPE	4798 (3.9%)
- WESTERN CAPE	21412 (17.2%)

A.2.1 Summary Data by Multiple Variables

To generate summary table for multiple variables, stratified by one variable, say HIV status, the code A.2 can be used.

```
table_two <- tableby(HIVStatusDerived ~
 SiteOfDiseaseDerived + PatientCaregoryDerived +
PreTreatmentMicrobiology + IsOnARVDerived + IsOnCPTDerived
,data = predictors_desc_stat)
clinical_xters <- summary(table_two, title = "predictors_
 desc_stat", text = TRUE, test = FALSE)
```

Code A.2: Code to derive summary statistics table for multiple predictos, stratified by HIV status.

	Positive (N=72321)	Negative (N=31049)	Unknown (N=21114)	Total (N=124484)
SiteOfDiseaseDerived				
- PTB	68370 (94.5%)	29663 (95.5%)	16185 (76.7%)	114218 (91.8%)
- EPTB	941 (1.3%)	396 (1.3%)	82 (0.4%)	1419 (1.1%)
- Unknown	3010 (4.2%)	990 (3.2%)	4847 (23.0%)	8847 (7.1%)
PatientCaregoryDerived				
- New	31904 (44.1%)	15623 (50.3%)	4823 (22.8%)	52350 (42.1%)
- Previously treated	39973 (55.3%)	15243 (49.1%)	12283 (58.2%)	67499 (54.2%)
- Unknown	444 (0.6%)	183 (0.6%)	4008 (19.0%)	4635 (3.7%)
PreTreatmentMicrobiology				
- Positive	61740 (85.4%)	26674 (85.9%)	6623 (31.4%)	95037 (76.3%)
- Negative	1810 (2.5%)	927 (3.0%)	2289 (10.8%)	5026 (4.0%)
- Unknown	8771 (12.1%)	3448 (11.1%)	12202 (57.8%)	24421 (19.6%)
IsOnARVDerived				
- Yes	63405 (87.7%)	0 (0.0%)	0 (0.0%)	63405 (50.9%)
- No	3677 (5.1%)	0 (0.0%)	0 (0.0%)	3677 (3.0%)
- Unknown	5239 (7.2%)	0 (0.0%)	0 (0.0%)	5239 (4.2%)
- NULL	0 (0.0%)	31049 (100.0%)	21114 (100.0%)	52163 (41.9%)
IsOnCPTDerived				
- Yes	42785 (59.2%)	0 (0.0%)	0 (0.0%)	42785 (34.4%)
- No	3609 (5.0%)	0 (0.0%)	0 (0.0%)	3609 (2.9%)
- Unknown	25927 (35.8%)	0 (0.0%)	0 (0.0%)	25927 (20.8%)
- NULL	0 (0.0%)	31049 (100.0%)	21114 (100.0%)	52163 (41.9%)

A.3 Mapping Cleaned ETR to EDRweb

Table A.1: Mapping cleaned ETR data to EDR database

Column in ETR	Column in EDR_DB	Table in EDR_DB	Comment
DateStamp	DateStamp	Facility	There is another column named DateStamp in Person table, lots of them are null. Person table can only be linked to episode table
TBRegistrationID	EpisodeRegistrationID	EpisodeRegistration	
IsActive	not applicable	not applicable	This is only applicable to ETR, there should be no equivalent in EDR
Province	ProvinceName	vwDHIS_Facility	contains all locations details rather than IDs in Facility table.Kindly check facility table
District	DistrictName	vwDHIS_Facility	same as above
SubDistrict	SubDistrictName	vwDHIS_Facility	same as above
Facility	Name	vwDHIS_Facility	Spelling seems consistent but some facilities appear in EDR or ETR only.
Facility_Active	IsActive	vwDHIS_Facility	also contained in Facility table, not sure if it is different from IsActive column in ETR
Facility_Decommisioned	IsDecommissioned	Facility	no problem on this one

...continued

Column in ETR	Column in EDR_DB	Table in EDR_DB	Comment
DateEntryLevel	NA	NA	Not applicable to EDR
Surname	Surname	Person	All null in ETR and EDR
FirstName	Name	Person	All null in ETR and EDR
Gender	Gender	Person	
Age	AgeAtTreatmentStart	Episode	age of patients at the start of treatment
DateOfBirth	BirthDate	Person	All null in EDR
IDNumber	IDNo	Person	All null in ETR and DR
RegistrationDate	RegistrationDate	Episode	most entries are null
RegistrationNo	RegistrationNo	EpisodeRegistration	
RegistrationYear	RegistrationYear	EpisodeRegistration	
Registration_Type	RegistrationTypeID	EpisodeRegistration	merge with look up table to extract the definition of IDs
TreatmentStartDate	TreatmentStartDate	Episode	
Treatmentyear	TreatmentStartYear	Episode	year part extracted from TreatmentStartDate
TB_Diagnosis_type	NA	NA	All subjects in EDR are confirmed TB
Regimen		NA	regimen is not applicable EDR as treatment is varied, list of drugs used is equivalent. Extract list of drugs used from SecondLineDrug table by systematically merging with Medication and Episode table
Patient_Category	PatientCategoryID	Episode	merge with look up table to extract the definition of IDs
IsTreatmentStarted	IsTreatmentStarted	Episode	1 represents Yes, 0 represent No. Extract meaning from Lookup table
Disease_Classification	SiteOfDiseaseID	Episode	merge with look up table to extract the definition of IDs
Site_Of_Disease_1	not found	not found	create this field the variable, SiteOfDisease. Site of disease 1 means pulmonary
Site_Of_Disease_2	not found	not found	create this field the variable, SiteOfDisease. Site of disease 2 means extra-pulmonary
Treatment_Outcome	OutcomeID	Episode	merge with look up table to extract the definition of IDs
OutcomeDate	OutcomeDate	Episode	
SystemGeneratedOutcome	Not found	not found	a field derived from the outcome field, by simply renaming entries and spcifically changing all null outcomes to 'Not evaluated'

...continued

Column in ETR	Column in EDR_DB	Table in EDR_DB	Comment
MovedFrom	SourceFacilityID	TransferNotification	match source Facility ID here to FacilityID in Facility table
MovedTo	DestinationFacilityID	TransferNotification	match desination Facility ID here with FacilityID in Facility table
TransferType	TranferTypeID	EpisodeRegistration	merge with look up table to extract the definition of IDs
TransferLocation	FromFacilityID	PatientTransfer	merge with facility table to extract the definition of IDs
Not found	ToFacilityID	PatientTransfer	merge with facility table to extract the definition of IDs. This field is an extra field, it is not available in ETR
Treatment_Outcome_Combined	not found	not found	this is a derived field that replaces missing outcomes with tranfer histrory, if available
SystemGeneratedOutcome_MT	not found	not found	a renamed field from the field above
PreTreatmentSputumType	not found	not found	a field derived from initial and baseline sputum (and culture) results. It is positive if any of sputum or culture result is positive
Month2_Sputum_Result	SputumResultID	PatientSputum	match with look up table to extract the definition of IDs and also specifically match with month 2 treatment month field in sputum table
Month3_Sputum_Result	SputumResultID	PatientSputum	match with look up table to extract the definition of IDs and also specifically match with month 3 treatment month field in sputum table
HIV_Status	HIVStatusID	Episode	merge with look up table to extract the definition of IDs
HIV_Test_Result	HIVResult	PatientHIVTest	
CD4Result	CD4Count	PatientHIVTest	also available as cd4 in BloodResults table
ART_Started	ARVStartedID	Episode	merge with look up table to extract the definition of IDs
CPT_Started	CPTSTartedID	Episode	merge with look up table to extract the definition of IDs
HIV_Status_Derived	not found	not found	redundant field
HIV_Status_Reason	not found	not found	a derived field from HIV_Status
TB_Diagnosis_Type_Derived			redundant field
TreatmentYear_Derived			redundant filed
Not found	TypeOfResistantTB	Episode	a field of resistance pattern
Not found	PersonID	Person	uniquely to each person

References

[1] WHO. *Global tuberculosis report 2019*. World Health Organization, 2019.

[2] National Institute for Communicable Diseases. *World TB Day*, 2019 (accessed November 23, 2019). URL http://www.nicd.ac.za/world-tb-day-2019.

[3] Alex Welte, Elisabetta Walters, and Robert Gie. Tuberculosis: well known for centuries - still difficult to diagnose. *SACEMA quarterly*, 2018.

[4] Hans L Rieder et al. *Epidemiologic basis of tuberculosis control*. Number Ed. 1. International Union Against Tuberculosis and Lung Disease (IUATLD), 1999.

[5] Centers for Disease Control and Prevention. *Latent TB Infection and TB Disease*, 2016 (accessed November 24, 2019). URL https://www.cdc.gov/tb/.

[6] P Glaziou, C Sismanidis, M Zignol, and K Floyd. Methods used by who to estimate the global burden of tb disease. *Global TB Programme, WHO, Geneva*, 2016.

[7] TB DOTS Strategy Coordination. *National TB Management Guidelines 2014*. National Department of Health, 2014.

[8] Tetiana Zhabska. *Tuberculosis is an infection caused by bacteria*, 2016 (accessed November 24, 2019). URL https://www.shutterstock.com/image-vector/tuberculosis-infection-caused-by-bacteria-infected-235957408.

[9] Michael Walsh. *Tuberculosis*, 2013 (accessed November 23, 2019). URL http://www.infectionlandscapes.org/2013/04/tuberculosis.html.

[10] World Health Organization et al. Guidance for national tuberculosis programmes on the management of tuberculosis in children. Technical report, World Health Organization, 2014.

[11] Annabel Kanabus. *Information about tuberculosis*, 2018 (accessed November 26, 2019). URL https://www.tbfacts.org.

[12] Chang Liu, Christopher J Lyon, Yang Bu, Zaian Deng, Elisabetta Walters, Yan Li, Liqun Zhang, Anneke C Hesseling, Edward A Graviss, and Ye Hu. Clinical evaluation of a blood assay to diagnose paucibacillary tuberculosis via bacterial antigens. *Clinical chemistry*, 64(5):791–800, 2018.

[13] Rashmi Kumar. Evaluation of diagnostic tests. *Clinical Epidemiology and Global Health*, 4(02):76–79, 2016.

[14] Karlijn J Van Stralen, Vianda S Stel, Johannes B Reitsma, Friedo W Dekker, Carmine Zoccali, and Kitty J Jager. Diagnostic methods i: sensitivity, specificity, and other measures of accuracy. *Kidney international*, 75(12):1257–1263, 2009.

[15] Yemisi Takwoingi and Terence J Quinn. Review of diagnostic test accuracy (DTA) studies in older people. *Age and ageing*, 47(3):349–355, 2018.

[16] Bruno Dujardin, Jef Van den Ende, Alfons Van Gompel, Jean-Pierre Unger, and Patrick Van der Stuyft. Likelihood ratios: a real improvement for clinical decision making? *European journal of epidemiology*, 10(1):29–36, 1994.

[17] Margaret Sullivan Pepe et al. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

[18] Andrew Owen. Bayes theorem and diagnostic tests with application to patients with suspected angina. In *International Cardiovascular Forum Journal*, volume 1, pages 96–100, 2013.

[19] Douglas M Hawkins, James A Garrett, and Betty Stephenson. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in medicine*, 20(13):1987–2001, 2001.

[20] P Finbarr Duggan. Time to abolish "gold standard". *BMJ: British Medical Journal*, 304(6841):1568, 1992.

[21] E Versi. "gold standard" is an appropriate term. *BMJ: British Medical Journal*, 305(6846):187, 1992.

[22] World Health Organization. *Information about tuberculosis*, 2019 (accessed November 26, 2019). URL https://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.

[23] Inge K Holden, Troels Lillebaek, Niels Seersholm, Peter H Andersen, Christian Wejse, and Isik S Johansen. Predictors for pulmonary tuberculosis treatment outcome in denmark 2009–2014. *Scientific reports*, 9(1):1–8, 2019.

[24] Ato Kwamena Tetteh, Edward Agyarko, Joseph Otchere, Langbong Bimi, and Irene Ayi. An evaluation of treatment outcomes in a cohort of clients on the dots strategy, 2012–2016. *Tuberculosis research and treatment*, 2018, 2018.

[25] Chunqin Long, Huaxiang Xu, Qinglin Shen, Xianghai Zhang, Bing Fan, Chuanhong Wang, Bingliang Zeng, Zicong Li, Xiaofen Li, and Honglu Li. Diagnosis of the coronavirus disease (covid-19): rrt-pcr or ct? *European journal of radiology*, page 108961, 2020.

[26] Victor M Corman, Olfert Landt, Marco Kaiser, Richard Molenkamp, Adam Meijer, Daniel KW Chu, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Marie Luisa Schmidt, et al. Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr. *Eurosurveillance*, 25(3):2000045, 2020.

[27] Savitskaya Judy and Conde Jorge. *Coronavirus is Testing the World. How Do We Test for Coronavirus?*, 2020 (accessed May 18, 2020). URL https://www.c-sharpcorner.com/uploadfile/nipuntomar/normalization-and-its-types/.

[28] S Jones, S Carley, and M Harrison. An introduction to power and sample size estimation. *Emergency medicine journal: EMJ*, 20(5):453, 2003.

[29] Khalid S Khan, Gerben Ter Riet, Julie Glanville, Amanda J Sowden, Jos Kleijnen, et al. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*. Number 4 (2n. NHS Centre for Reviews and Dissemination, 2001.

[30] Anne WS Rutjes, Johannes B Reitsma, Jan P Vandenbroucke, Afina S Glas, and Patrick MM Bossuyt. Case–control and two-gate designs in diagnostic accuracy studies, 2005.

[31] Mariska MG Leeflang and Franz Allerberger. Sample size calculations for diagnostic studies. *Clinical Microbiology and Infection*, 25(7):777–778, 2019.

[32] Karimollah Hajian-Tilaki. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, 48:193–204, 2014.

[33] Shein-Chung Chow, Jun Shao, Hansheng Wang, and Yuliya Lokhnygina. *Sample size calculations in clinical research*. Chapman and Hall/CRC, 2017.

[34] Gerald Van Belle, Lloyd D Fisher, Patrick J Heagerty, and Thomas Lumley. *Biostatistics: a methodology for the health sciences*, volume 519. John Wiley & Sons, 2004.

[35] Kenneth J Rothman and Sander Greenland. Planning study size based on precision rather than power. *Epidemiology*, 29(5):599–603, 2018.

[36] Jun-Mo Nam. Optimum sample sizes for the comparison of the control and treatment. *Biometrics*, pages 101–108, 1973.

[37] William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

[38] Mitchell Gail, Roger Williams, David P Byar, and Charles Brown. How many controls? *Journal of chronic diseases*, 29(11):723–731, 1976.

[39] Jia Fan, Hedong Zhang, Duc T Nguyen, Christopher J Lyon, Charles D Mitchell, Zhen Zhao, Edward A Graviss, and Ye Hu. Rapid diagnosis of new and relapse tuberculosis by quantification of a circulating antigen in hiv-infected adults in the greater houston metropolitan area. *BMC medicine*, 15(1): 188, 2017.

[40] Chang Liu, Zhen Zhao, Jia Fan, Christopher J Lyon, Hung-Jen Wu, Dobrin Nedelkov, Adrian M Zelazny, Kenneth N Olivier, Lisa H Cazares, Steven M Holland, et al. Quantification of circulating mycobacterium tuberculosis antigen peptides allows rapid diagnosis of active disease and treatment monitoring. *Proceedings of the National Academy of Sciences*, 114(15):3969–3974, 2017.

[41] Edmond De Hoffmann. Mass spectrometry. *Kirk-Othmer Encyclopedia of Chemical Technology*, 2000.

[42] Leigh Anderson and Christie L Hunter. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Molecular & Cellular Proteomics*, 5(4):573–588, 2006.

[43] Keiji Kito and Takashi Ito. Mass spectrometry-based approaches toward absolute quantitative proteomics. *Current genomics*, 9(4):263–274, 2008.

[44] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[45] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[46] Ben Van Calster, Maarten van Smeden, Bavo De Cock, and Ewout W Steyerberg. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, 2020.

[47] Nipun Tomar. *Normalization and its Types*, 2009 (accessed June 20, 2020). URL https://www.c-sharpcorner.com/uploadfile/nipuntomar/normalization-and-its-types/.

[48] Directorate Drug-Resistant TB. Multi-drug resistant tuberculosis: A policy framework on decentralised and deinstitutionalised management for south africa. *Infection Control. Directorate Drug-Resistant TB, Pretoria*, 2019.

[49] World Health Organization. *WHO consolidated guidelines on drug-resistant tuberculosis treatment*. World Health Organization, 2019.

[50] Ethan Heinzen, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson, and Gregory Dougherty. *arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries*, 2019. URL https://CRAN.R-project.org/package=arsenal. R package version 3.3.0.

[51] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL https://CRAN.R-project.org/package=xtable. R package version 1.8-4.